

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
31 October 2002 (31.10.2002)

PCT

(10) International Publication Number  
**WO 02/086163 A1**

- (51) International Patent Classification<sup>7</sup>: **C12Q 1/68**
- (21) International Application Number: **PCT/SE02/00788**
- (22) International Filing Date: **22 April 2002 (22.04.2002)**
- (25) Filing Language: **English**
- (26) Publication Language: **English**
- (30) Priority Data:  
**60/284,925** **20 April 2001 (20.04.2001)** **US**
- (71) Applicant (*for all designated States except US*):  
**KAROLINSKA INNOVATIONS AB [SE/SE]; S-171 77 Stockholm (SE).**
- (72) Inventors; and
- (75) Inventors/Applicants (*for US only*): **ZABAROVSKY, Eugene [RU/SE]; Åvägen 7, S-194 44 Upplands Väsby (SE). ERNBERG, Ingemar [SE/SE]; Smultrongränd 18, S-182 45 Enchyberg (SE). LI, Jingfeng [CN/SE]; Emmylundsvägen 3, 346, S-171 72 Solna (SE). PRO-TOPOPOV, Alexei [RU/SE]; Professorslingan 41, 204, S-104 05 Stockholm (SE). VORONTSOVA, Olga [UA/SE]; Professorslingan 41, 202, S-104 05 Stockholm (SE). WAHLESTEDT, Claes [SE/SE]; Strandvägen 29, S-114 56 Stockholm (SE). KASHUBA, Vladimir [UA/SE]; Rinkebyvägen 87/201, S-163 74 Spanga (SE). ZABAROVSKA, Veronika [RU/SE]; Åvägen 7, SE-194 44 Upplands-Väsby (SE).**
- (74) Agent: **DR LUDWIG BRANN PATENTBYRÅ AB; Drottninggatan 7, Box 1344, S-751 43 Uppsala (SE).**
- (81) Designated States (*national*): AE, AG, AL, AM, AT (utility model), AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ (utility model), CZ, DE (utility model), DE, DK (utility model), DK, DM, DZ, EC, EE (utility model), EE, ES, FI (utility model), FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK (utility model), SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.
- (84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).
- Published:
- *with international search report*
  - *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*



**WO 02/086163 A1**

(54) Title: **METHODS FOR HIGH THROUGHPUT GENOME ANALYSIS USING RESTRICTION SITE TAGGED MICROARRAYS**

(57) Abstract: The present invention provides novel and unique techniques for high-throughput analysis of genomic material originating from complex biological systems, including complex microbial systems. The present invention also pertains to a method of detecting changes in a genomic material using restriction site tagged (RST) microarrays and sequence passporting technique (in particular microarrays containing NotI-clones). Using the present invention method, methylation or silencing of specific alleles, homozygous and hemizygous deletions, epigenetic factors, genetic predisposition, etc, information which is particularly useful in diagnosis and treatment of cancer diseases, can be detected. The RST microarrays and passporting according to the present invention can also be used for qualitative and quantitative analysis of complex microbial systems.

## Methods for high throughput genome analysis using restriction site tagged microarrays.

### 5 FIELD OF THE INVENTION

The present invention pertains to a method of detecting changes in a genomic material using restriction site tagged (RST) microarrays and passporting technique, which can be used for detecting methylation or silencing of specific alleles, homozygous, hemizygous deletions, epigenetic factors, genetic predisposition, etc, information which  
10 is particularly useful in diagnosis and treatment of cancer diseases. The RST microarrays and passporting according to the present invention can also be used for qualitative and quantitative analysis of complex microbial systems.

### BACKGROUND OF THE INVENTION

15 Genomic subtractive methods in principle are very useful for identification of disease genes including tumour suppressor genes. However, among many suggested techniques only a modified variant of genomic subtraction called Representational Difference Analysis (RDA, Lisitsyn et al., 1993) and RFLP subtraction (Restriction Fragment Length Polymorphism)(Rosenberg et al., 1994) have been reproducibly  
20 succesful in cloning deleted sequences. Three main drawbacks limited wide use of these related methods: both are very complicated and laborious, they are very sensitive to minor impurities and experiments result in cloning only a few deleted sequences. It is important to note that these methods only work well with enzymes not being associated with CpG islands. Methylation-sensitive-representational analysis (MS-RDA,  
25 Ushijima et al., 1997) has more specific aims, i.e. they work with CpG Islands, but still is not avoided limitations of the original RDA. Moreover, differentially cloned products usually do not have any connections with genes. Deletions of non-functional regions occur frequently in the human genome and cloning of such segments will not yield valuable information (Lisitsyn et al., 1995). RDA is also unable to detect differences  
30 due to point mutations, small deletions or insertions, unless they affect a particular restriction enzyme recognition site. Another source of artefacts is the PCR amplification after the first hybridization step and before the nuclease treatment. The presence of excess driver DNA can result in a reduced efficiency of the amplification

- tester:tester duplexes due to the opportunity for the residual driver:driver and driver:tester duplexes to act as competitors. As RDA is based mainly on specific PCR amplification of desired products and use many cycles (95-110), it suffers from a "plateau effect" that is characterised by a decline in the exponential rate of accumulation of amplification products (Innins and Gelfand, 1990). However, the major problem results from the inefficiency of the multiple restriction digestion and ligation reactions that are used in this method and leads to the generation of false positives.
- 10 The presence of genetic alterations in tumours is now widely accepted, and explains the irreversible nature of tumours. However, observations on tissue differentiation indicated that it shares something in common with carcinogenesis, i.e. "epigenetic" changes. Now, DNA methylation in CpG sites is known to be precisely regulated in tissue differentiation, and is supposed to be playing a key role in the control of gene expression in mammalian cells. The enzyme involved in this process is DNA methyltransferase, which catalyzes the transfer of a methyl group from S-adenosylmethionine to cytosine residues to form 5-methylcytosine, a modified base that is found mostly at CpG sites in the genome. The presence of methylated CpG islands in the promoter region of genes can suppress their expression. This process may be due to the presence of 5-methylcytosine that apparently interferes with the binding of transcription factors or other DNA-binding proteins to block transcription. DNA methylation is connected to histone deacetylation and chromatin structure, and regulatory enzymes of DNA methylation are being cloned.
- 25 In different types of tumours, aberrant or accidental methylation of CpG islands in the promoter region has been observed for many cancer-related genes resulting in the silencing of their expression. The genes involved include tumour suppressor genes, genes that suppress metastasis and angiogenesis, and genes that repair DNA, suggesting that epigenetics plays an important role in tumorigenesis. The potent and specific inhibitor of DNA methylation, 5-aza-2-deoxycytidine (5-AZA-CdR) has been demonstrated to reactivate the expression of most of these malignant suppressor genes in human tumour cell lines. These genes may be interesting targets for chemotherapy with inhibitors of DNA methylation in patients with cancer, and may help to clarify the
- 30

importance of this epigenetic mechanism in tumourigenesis. Spontaneous regression of malignant tumours used to enchant researchers, but it has now been observed that genes inactivated by hypermethylation are frequently involved in tumours that relatively often undergo spontaneous regression. Carcinogenic mechanisms of some  
5 carcinogens seem to involve modifications of an epigenetic switch, and some dietary factors also have the possibility to modify the switches.

Review articles in the literature make it clear that methylation is a basic, vital feature/mechanism in mammalian cells. It is involved in hereditary and somatic  
10 cancers, hereditary and somatic diseases, apoptosis, replication, recombination, temperature control, immune response, mutation rate (i.e. in p53). Through methylation food can induce cancer, etc., it is believed that it can be used for diagnostic, prognostic, prediction and even for direct treatment of cancer. Inactivation of DNA methyltransferase is lethal for mice. Based on the growing understanding of  
15 the roles of DNA methylation, several new methodologies have been developed to make a genome-wide search for changes in DNA methylation.

There are four main genome-wide screening methods (see Sugimura T, Ushijima T, 2000) for testing methylation in human genome: restriction landmark genomic  
20 scanning (RLGS, Costello et al., 2000), methylation-sensitive-representational difference analysis (MS-RDA), methylation-specific AP-PCR (MS-AP-PCR) and methyl-CpG binding domain column/segregation of partly melted molecules (MBD/SPM). Although each of them has their own advantages, none of them is suited for large-scale screening since all four are rather inefficient and complicated; they can be used only  
25 for testing a few samples. For example, after analysis of 1000 clones isolated using MBD/SPM, nine DNA fragments were identified as CpG islands and only one was specifically methylated in tumour DNA.

Recently developed microarrays of immobilized DNA open new possibilities in  
30 molecular biology. These DNA arrays, containing either cDNA or genomic DNA, are fabricated by high speed robotics on glass substrates. Probes that are labeled by different colors are hybridized. In one such hybridization thousands of genes or genomic DNA fragments can be analyzed allowing massive parallel gene expression and

gene discovery studies. In pilot experiments microarrays with immobilized P1 and BAC clones DNA demonstrated that they could be used for high resolution analysis of DNA copy number variation using CGH (comparative genome hybridization). It has been suggested that this approach can work if inserts of human DNA in the cloning vectors are larger than 50 kb. In the future, when microarrays with P1 and BAC clones covering the whole human genome will be created, this approach will most likely replace conventional CGH. Clearly, construction of such microarrays with mapped P1 and BAC clones is very expensive, laborious and time consuming. Construction of such microarrays cannot be achieved in a single research laboratory. If small-insert NotI linking clones could fulfill the same function this will open the way to construct such microarrays for CGH analysis for a single research group and for many organisms. PACs and BACs covering the whole human genome are not available yet.

Pollack et al., 1999 suggested to use cDNA microarrays for genomic DNA copy number changes but small size of cDNA clones and high ratio of background hybridization compared to real signal makes this suggestion problematic.

In the fall 2000 Affymetrix launched the selling of GeneChipHuSNP Mapping Assay. These microarrays contain 1.494 SNP loci. In the promotion papers it was shown that this microarrays can be used for the detection of loss of heterozygosity (LOH). However 13% of SNPs failed in the majority of samples whereas only 354 SNPs were informative in one particular experiment.

Lucito et al. (2000) used for the detecting copy number fluctuations in tumour cells modification of RDA technology. In this method BglII representations were used in conjunction with DNA microarrays. As there are many small BglII clones in the human genome (150.000) it will be not easy and cheap to make comprehensive microarrays with unique clones covering the whole human genome.

Presently, there are some methods available to analyze complex microbial mixtures, e.g. by enzyme analysis (Katouli et al., 1994) which requires growth of colonies outside the body, or analysis of the composition fatty acids in stools which gives crude indications of the composition of the normal flora (refs.), however all them have

obvious limitations.

The application of culture-independent techniques based on molecular biology methods that can overcome some shortcomings of conventional cultivation methods. In recent years the approaches based on PCR amplification of 16S rRNA genes have been most popular. One modification of the approach utilized fingerprinting of all the species in the gut using, for instance, denaturing gradient gel electrophoresis (DGGE) with PCR amplified fragments of 16S rRNA genes. In another application, PCR amplified fragments of 16S rRNA genes were directly cloned and sequenced. These studies yielded important information however intrinsic disadvantage of the approach limits its application. The problem is that 16S rRNA genes are highly conserved and therefore the same sequenced fragment can belong to different species. It is also important to keep in mind that in fingerprinting experiments similar fragments can represent different species, and different fragments can represent the same species.

#### SUMMARY OF THE INVENTION

In view of the drawbacks associated with the prior art methods for analysis of genomic material originating from complex biological systems, there is a need for uncomplicated, quick and reliable genome analysis methods.

Therefore, the object of the present invention is to provide novel and unique techniques for analysis of genomic material originating from complex biological systems, including complex microbial systems. The main objects of the present invention are the following:

One object of the present invention is to prepare and to use NotI-clone (in general PCR fragments, oligonucleotides, etc.) microarrays for studying methylation and/or copy number changes in eukaryotic genomes for diagnosis, prognosis, identification of cancer causing genes. NotI microarrays are the only existing microarrays giving the opportunity to detect copy number changes and methylation simultaneously. This includes comparison of normal and malignant cells at genomic and/or RNA level; comparison of primary tumours and metastases; analysis of families suffering from hereditary diseases including cancers; and diagnostics and disease prediction.

Capability to establish differences between normal and tumour cells is instrumental for cloning cancer causing genes and for early diagnosis and prevention of cancer. It is also very important for differentiation, development and evolution studies.

- 5 Another object of the present invention is to provide techniques allowing qualitative and quantitative analysis of complex microbial systems, such as the normal flora of the gut.

A further object of the present invention is to prepare NotI sequencing passports ("NotI passport") (collection of NotI tags: short sequences surrounding genomic NotI sites)  
10 and to use them to study the same problems as were mentioned above for NotI microarrays.

Wide screening of genomic material using RST encounter many problems, e.g. the size  
15 of the human genome/microbial mixture and the number of repeat sequences. We have solved these problems by developing a new method for labeling genomic DNA, where only sequences surrounding NotI (or any other restriction) sites are labeled (tagged), herein called NotI Representation (NR).

20 In the present invention, Restriction Site Tags (RSTs) are generated from thousands of microorganisms or human genomes and used for the generation of NotI RST microarrays passports which describe uniquely not only individual human cell/organism or bacterial strains but most or all the members of a microbial flora of e.g. in the gut.

25

With the NotI or RST genome scanning method according to the present invention, large scale scanning of microbial genomes on a quantitative and qualitative basis is possible.

30 From the results of our experiments, we have shown that it is possible to create a large database containing NotI microarrays passports, i.e. NotI microarray images. Many samples of colon flora have been compared to determine their exact composition.

The present invention procedure is universal, i.e. we can use any other enzyme for creating "RST microarray passports". Moreover, any biochemical or chemical approach cutting DNA (RNA) in a specific position scarcely distributed along DNA (RNA) can be used. For example, it can be enzyme like cre-recombinase or chemically modified  
5 oligonucleotide forming triplex DNA and initiating DNA break. The polymorphism of NotI representations can be increased by using several enzymes in addition to BamHI, e.g. BclI, BglII, HindIII etc. In pilot experiments we have produced NotI microarrays from gram-positive and gram-negative bacteria and have shown that even very similar E. coli strains can be easily discriminated using this technique. Using the above  
10 mentioned technique we can identify important pathogenic bacteria in the human organism.

These 'NotI microarrays passports' can be produced for individuals, normal/tumour pairs, different cell NotI Representation (NR). A pilot experiment using NR probes  
15 demonstrated the power of the method, and we successfully detected Chr.3 NotI clones deleted in ACC-LC5 and MCH939.2 cell lines.

Such NotI RST microarrays can be prepared for any human or any groups of humans, who for example suffer from the same specific disease, in order to detect a certain  
20 disease which cannot be detected by other means. NotI RST microarrays can also be prepared for any mammal (like cattles or dogs) or microbial organism.

NotI arrays will speed up cancer research very significantly and can replace CGH, LOH and many cytogenetic studies.

25

The NotI scanning approach will find mainly deleted, amplified, or methylated genes but it will also identify polymorphic and mutated NotI sites. Comparing these NotI passports can give a clue to understanding many diseases and other fundamental biological processes.

30

Using the present invention method of producing RST microarrays, restriction enzyme tagged (RST) microarrays for any enzyme can be created. The microarrays according to the present invention represent a novel type of microarrays, which is completely



different from the existing ones (oligonucleotides, cDNA, genomic BAC/PAC clones).

To be able to establish differences between individual compositions of the normal gut flora will be instrumental for future analysis of how the normal flora composition is influenced by diet, special foods, geographical location, colon, ovarian, etc. cancers and other diseases. It has particularly wide applications for cancer research.

The present invention method will probably have strong impact both on basic science and on human and animal health, agriculture, medicine, pharmacology, etc.

10

We propose to use our NotI clones as a complement to microarrays based on P1 and BAC clones covering the whole human genome. Microarrays based on small-insert NotI linking clones have been developed, and can have a similar function. Approximately 10.000-20.000 NotI clones, covering the whole human genome and containing 10%-20% of all genes (40%-50% of them are not present in ESTs microarrays) are already available.

15

In order to achieve what is described above, the present invention comprises the following embodiments:

20

In one embodiment of the present invention provides a method for preparing nucleic acid or and/or modified nucleic acid reference material bound to a solid phase, comprising the steps of

-digesting nucleic acid and/or modified nucleic acid reference material using biochemical and/or chemical approaches, to obtain sequence fragments surrounding a specific recognition site,  
-selecting said nucleic acid and/or modified nucleic acid sequence fragments associated with a specific recognition site.

25

Said reference material is digested by a first restriction enzyme and/or one or more second restriction enzymes, e.g. endonucleases, such as cre-recombinase,

30

In one embodiment of the present invention the recognition sites of the first

endonuclease is scarcely distributed along said genomic material and is located adjacent to gene sequences, and the recognition sites of said one or more second restriction endonucleases are more frequently occurring along said genomic material than the sites of the first endonuclease.

5

In another embodiment of the present invention the digestion by the first and second restriction endonucleases are performed simultaneously, and different linkers are ligated to the ends resulting from cutting by the first and second restriction endonucleases, respectively, which linkers are designed such that when primers are added in order to make PCR reactions, only the fragments containing ends resulting from cutting by the first restriction endonuclease will be amplified.

10

In still another embodiment of the present invention the reference material is first digested by the one or more second restriction endonucleases, the ends of the thus obtained fragments are self-ligated into the form of circular nucleic acid and/or modified nucleic acid molecules, and any linear fragments remaining after self-ligation are inactivated before digestion with the first restriction endonuclease, whereby the linear fragments resulting from the digestion by the first endonuclease are subjected to PCR amplification.

15

In these embodiments the first restriction endonuclease is NotI, or any other restriction endonuclease, the restriction sites of which occurs in proximity to CpG islands in the genomic material.

20

The first restriction endonuclease can also be NotI, PmeI or SbfI, or a combination of two or more of said endonucleases, and the second endonuclease can be BamHI, BclI, BglII or Sau3A, or a combination of two or more of said endonucleases.

Said nucleic acid and/or modified nucleic acid reference material can be selected from RNA, DNA, peptides or modified oligonucleotides, or a combination of two or more of said materials.

25

In the present invention nucleic acid and/or modified nucleic acid is bound to a solid

30

glass support in the form of a microarray. However, the present invention is not limited to using glass microarrays. Solid phases such as filters, e.g. nylon filters, coded beads, cellulose, such as nitrocellulose, or other solid supports can also be used to bind nucleic acid and/or modified nucleic acid. In general DNA, oligonucleotides, etc. bound to a solid phase can be used.

The genomic material that can be used according to the present invention can be derived from one or more humans, from different locations in the body/bodies and at the same or different points in time. Said genomic material can be derived from bacteria from the gut, skin or other parts of the human body. However, it can also be derived from any organism, bacteria, animal, or plant, or product produced therefrom, or from any substance wherein genomic material can be contained, especially air and water.

The present invention also pertains to the fragments that can be obtained using the present invention, and the nucleic acid or and/or modified nucleic acid microarrays containing these fragments.

The present invention further pertains to representations of the genome, or of a part thereof, of an organism, comprising multiple copies of the nucleic acid and/or modified nucleic acid fragments, or a selection thereof, obtained by means of the present invention method.

These representations, in liquid form, are hybridized to the nucleic acid and/or modified nucleic acid fragments present in the form of said solid phases.

Said representations can be used for discriminating between different genomes, detecting methylations, deletions, mutations and other changes within genomic material obtained from the same individual at different points of time, or in the genomic material obtained from one individual as compared to a standard representation obtained from at least one other individual, or a combination thereof.

In addition to the above-mentioned applications, these representations can be used

for:

- studying methylation and copy number changes in eukaryotic genomes for diagnosis, prognosis, identification of cancer causing genes, etc,
- genotyping different microorganisms (viruses, prokaryotic, eukaryotic),
- 5 -studying biocomplexity and diversity of complex biological systems, i.e. human gut, bacterial flora in water, food, air resources,
- identifying pathogenic organisms in different sources including complex biological mixtures,
- producing passports (images of microarrays hybridizations, databases containing tag sequences) for different purposes: to describe organisms at
- 10 different conditions, i.e. different ages, disease/healthy, infected/uninfected etc,
- identifying new organisms, e.g. bacterial species,
- producing microarrays (DNA- and oligo-based) to study all above
- 15 described features,
- verification and maintenance of large biological collection/banks, i.e. verifying cell lines and individual organisms for higher organisms and confirming the purity of the particular strain for microbial species,
- producing kits for labeling and hybridization with microarrays,
- 20 -producing kits for making sequence tagging (passporting), and
- producing oligo microarrays to analyze sequence tags,

Finally, the present invention also pertains to a NotI CODE genomic subtraction method based on the use of the above described fragments.

25

#### BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1. General scheme for the NotI-CODE subtractive procedure.

Figure 2. Southern hybridization of NotI clones showed different hybridization. Clone names are shown at the bottom. N – normal DNA, L – DNA isolated from lung cancer cell line ACC-LC5.

30

Figure 3. General principle of using NR for NotI microarrays.

Figure 4. *NotI* microarrays profiling of deletions/methylation in microcell hybrid MCH 939.2 (A), cell line ACC-LC5 (B), and primary RCC tumors #196 (C) and #301 (D).

Representative images of microarrays (1) are ordered according to physical map of chromosome 3. One-dimensional clustering (2) is based on average normalized red/green ratios of fluorescent data (red,  $R > 3$ ; green,  $R < 0.3$ ). For (A) and (B) normal and tested DNA were hybridized together. NR for MCH903.1 (the whole chromosome) was labeled red and NR for MCH939.2 (3p.14-p22 deletion) was labeled green. Similarly, NR for normal lymphocyte DNA was red and small cell lung cancer line ACC-LC5 was labeled green. The red clusters demonstrate a significant overrepresentation of complete chromosome 3 or normal DNA. The green clusters – under representation of normal DNA. For (C) and (D) one step of *NotI*-CODE subtraction procedure was performed and single color hybridization was done. The green clusters demonstrate the significant overrepresentation of normal DNA. Grey color marks controls.

Figure 5. General scheme of the experiment. (microbial flora)

Figure 6. Flow chart diagram explaining generation of 85 bp oligonucleotide containing information about 19 bp *NotI*-tag

15

#### DETAILED DESCRIPTION OF THE INVENTION

In the literature it has been suggested and demonstrated that *NotI* sites are practically exclusively located in CpG islands and are closely associated with functional genes. Thus *NotI* sites are very useful markers not only for physical but also for genetic mapping.

20

The present inventors have created high-density grids that contain 50.000 of *NotI* clones originating from 6 representative *NotI* linking libraries and generated more than 22.000 unique *NotI* sequences (with stringent criteria 16.000) containing 17 Mb information. Analysis of these sequences demonstrated that even short sequences surrounding *NotI* sites is a source of important information allowing efficient isolation of new genes and the study of carcinogenesis.

25

We have a developed new approach for constructing *NotI* linking libraries (Zabarovsky et al., 1990) that give possibility to generate representative *NotI* linking libraries both in lambda phage and in plasmid form (Zabarovsky et al., 1994a). Since the procedure is quite easy and reproducible, it is possible to construct libraries from many sources.

30

Using the present invention NotI (RST) microarrays, based on the short sequences surrounding NotI sites or in general on restriction site tagged sequences (RSTS), complex biological systems, including complex microbial mixtures, can be qualitatively and quantitatively analysed.

5

In the present invention study NotI microarrays for human chr.3 (150 clones) were established and employed to compare chr 3 renal, lung, breast and nasopharyngeal cancers.

10 NotI microarrays for genome wide scanning

Recently we have sequenced 25.000 NotI clones and identified among them 16.000 unique clones. These clones that cover the whole human genome and contain 10%-20% of all genes (40%-50% of them are not present in ESTs microarrays) are already available.

15

The NotI microarrays can be used for testing tumour genomic DNA in genome wide NotI scanning (e.g. for deletion/amplification studies). Such arrays will speed up cancer research very significantly and can replace LOH (loss of heterozygosity), CGH (comparative genome hybridization), and other cytogenetic studies.

20

The fundamental problems for genome wide screening using NotI clones are:

- (i) the size and complexity of the human genome;
- (ii) the number of repeat sequences; and
- (iii) the comparatively small size of the inserts in NotI clones (on average 6-  
25 8 kb).

25

To solve this problem, the special primers were designed and special procedure was developed to amplify only regions surrounding NotI sites, so called NotI representation (NR). Other DNA fragments were not amplified. We suggested to use NotI microarrays  
30 for genome screening in combination with this new method for labeling genomic DNA where only sequences surrounding NotI sites are labeled.

NotI microarrays images can be generated for particular cells, tumours, and

individuals. By comparing images from normal and tumour cells, the differences between them will be defined. Using this information, NotI linking clones will be identified that differ between two (or more) DNAs. These clones can be used for further analysis and for isolating complete genes. Polymorphism in NotI sites is very frequent and according to the literature 43.5% of NotI sites are differently methylated or polymorphic.

Analysis of our database of 16.000 unique NotI sequences (two sequences can belong to the same NotI clone) showed that practically all of them are connected with genes and located at the 5' end of the genes. Comparison with completely sequenced chr. 21 and 22 revealed interesting observations. Chr. 21 contains 122 NotI sites (methylated and unmethylated) and Ichikawa et al., 1993 have cloned 40 NotI sites to construct the complete NotI restriction map with 43 NotI fragments. From these 40 clones our database contained 38 (95%) and additional 13 NotI clones (11%). Therefore using random sequencing we could isolate 27.5% more NotI clones than in the study of Ichikawa et al., 1993 where they focused their efforts in cloning NotI clones only from chr. 21. Altogether, from 390 possible NotI sites in chr. 21 and 22 our database contain 163 (42%) clones. Moreover, 18 clones that were identified in our work (5%) were not present in public sequences. These clones contained polymorphic NotI sites. Thus, from our data we can conclude that unmethylated (our database contain only unmethylated NotI sites) NotI sites represent appr. 42% and polymorphic ~ 5% of all possible NotI sites. Our estimation is that human genome contains 15.000-20.000 NotI sites and 6.000-9.000 of them are unmethylated in a particular cell. Thus screening with NotI microarrays will be equivalent to screening using 6.000-9.000 gene associated single nucleotide polymorphisms (SNP).

Comparing the prior art genomic chips with the present invention NotI microarrays it is easy to see that NotI microarrays give additional information to the deletion mapping: they can be used for gene expression profiling and methylation studies (see Table 1).

For preparing the probe for SNP chip 3.000 PCR primers and 24 separate reactions are needed and probe for NotI microarrays is prepared using 1-2 primers in one reaction

tube. Using the same NotI clones we are able to simultaneously obtain information about:

- (i) deletions/amplifications;
- (ii) methylation;
- 5 (iii) gene expression profiles.

All these features of NotI microarrays are extremely important for large scale experiments.

- 10 The pattern of hybridization of NR to the NotI microarrays represent a microarray passport for the DNA used for preparing NR.

We will now summarize the differences between CpG islands microarrays (below abbreviated to CGI, see Yan et al., Cancer Res. (2001) 61: 8375-8380), which we  
15 presently find is the closest prior art, and the present invention RST microarrays (below abbreviated to RST, see Table 2).

- 20 In the present invention sequences surrounding the same restriction site are cloned, whereas in CGI sequences originate from sequences between two restriction sites.

In principle, using the present invention technique, any restriction enzyme can be used for RST, but only limited number for CGI.

- 25 CGI can detect methylation, but not (in general) deletions (hemi- or homozygous) or amplifications of unmethylated sequences. RST can detect both copy number changes and methylation. CGI can detect deletion of the allele if it is methylated in normal genomic material and if it is deleted (unmethylated) in tumour material, this process is however inefficient as the vast majority of the important genes are unmethylated in  
30 normal genomic material, and the majority of methylated genes in normal genomic material are various kinds of repetitive elements, e.g. LINE, Long Interspersed Element (or sequence or repeat).



In CGI the total human DNA is labeled, in RST only 0.1-0.5%, and this DNA contains 10-fold less repeats than the total human DNA.

- 5 Many clones in CGI contain repeats and ribosomal DNA, whereas the RST only comprise genes containing unique human sequences. This very important difference is the result of completely different techniques of constructing microarrays (they use methyl-CG binding column, which is not used in the present invention).
- 10 For RST microarrays short OLIGOS (oligonucleotides 20-100 bp) can be used, which is not possible for CGI.

Incomplete digestion do not create problems for RST, but produce artificial signals in CGI.

15

Using RST hybridization is obtained when the site is not methylated, whereas in CGI hybridization only occurs if it is methylated.

- CGI microarrays can only be used to study methylation in high vertebrates. This can
- 20 also be done with RST, which in addition to that, also can be used for genotyping (passporting) any organism. It means that RST microarrays can be used to genotype bacteria and viruses for example, but not CGI.

- Our RST application contains complementary aspects, i.e. the generation of NotI (RST)
- 25 tags (passports) by sequencing. Sequencing can be done using different techniques including sequencing by hybridization to microarrays. No such complementary approach is possible with CGI.

- NotI-CODE (or RST-CODE in general) can be used together with RST microarrays to
- 30 remove in one step contaminating sequences. No such technique can be applied for CGI. Existing subtractive procedures like RDA cannot be employed, since they are not efficient enough to deal with the high complexity of total human genomic DNA.

Using RST microarrays it is possible to discriminate between deleted/amplified and methylated sequences. To achieve this aim NR should be produced using DNA that is unmethylated (it can be done by different approaches: limited PCR amplification after first digestion with restriction enzyme(s), enzymatic demethylation, etc.).

5

#### NotI passporting

We originally planned to use SAGE technique for this purpose. Serial analysis of gene expression (SAGE) allows for both a representative and comprehensive differential gene expression profile (Velculescu et al., 1995). The idea of the approach is that for each of  
10 the mRNA molecule a short 9-bp sequence tag is produced (including recognition site for the tagging enzyme it is 13 bp). Then these tags are ligated into concatemers and cloned. One sequencing reaction produces information for tens of RNA molecules. Thus by sequencing a few thousands clones one can e.g. evaluate all of the estimated 10.000 to 50.000 expressed genes in a given cell population. We have tried the SAGE  
15 technique for producing NotI tags but this was unsuccessful. Complexity of genomic DNA in microbial mixtures is at least 100 times more complex than the complexity of mRNA in eukaryotic cells. All RNA molecules must be tagged in SAGE but in our case, approximately one out of 250 molecules should be tagged. We propose to produce one tag for each 100-1.000 kb, but in SAGE one tag is produced for 256 bp. At the same  
20 time, a 13 bp tag is not enough for unambiguous identification of sequences in genomic DNA. That is why we have developed a new procedure called NotI passporting.

In this work we used the following modification. Genomic DNA was digested with NotI and ligated to the linker with NotI sticky ends. This linker contained BpmI recognition  
25 sites. This restriction nuclease cut 16/14 bp outside of the recognition site. Ligation mixture was digested with this enzyme to generate 11/9 nucleotide tags adjacent to the NotI site. This DNA sample was ligated to ZNBpm linker and PCR amplified with antiuniver and Z1univer primers to generate 85 bp duplex. The final PCR amplified molecule contains 17 bp sequence tag which is missing 2 bp from the original NotI site  
30 and therefore the whole NotI tag contains 19 bp. NotI passports were experimentally produced for E. coli K12, E. cloacae R4 and K. pneumoniae B4958. Experiments with samples obtained from mice demonstrated that the quality of DNA isolated from intestine of feces was sufficient to obtain NotI tags. The NotI passports uniquely

identified these species and among 96 tags none was common for these 3 bacterial species. Of course, ditags or concatemers also can be created from these 85 bp products. We believe that new high-throughput technologies like MPSS will make sequencing of single tags more efficient approach than creation of concatemers.

5 However, the design of the experiments can be different in different laboratories. As we mentioned above, this restriction site tagging procedure can be adapted to any recognition site for restriction nuclease. For comprehensive analysis of flora composition, use of several passports will be advantageous: different bacteria possess very different CG content. It means that with NotI passports bacteria having high CG  
10 content (NotI recognition site: GCGGCCGC) will predominantly be represented, but using for example SmaI passports (SmaI: ATTTAAAT), bacterial genomes with high AT content will be analyzed more carefully. Use of 2-3 different passports can significantly increase the sensitivity of the analysis and also be favourable for different applications, e.g. cancer risk, medication, diet, etc.

15 We tested the potentiality of the passporting approach and analyzed 25 bacterial species that were completely sequenced. The number of recognition sites for rare cutting restriction enzymes in these bacterial species are given in Table 3 below. It is easy to see that all 25 microbial species have different number of NotI recognition sites  
20 and therefore can be distinguished by NotI passporting. Moreover, from the Table 3 we can see that PmeI and SbfI restriction enzymes were even more informative.

Table 4 showed results of comparisons of different strains of *E. coli* and *Helicobacter pylori* for NotI, PmeI and SbfI enzymes. All of these strains were uniquely described by  
25 any of these enzymes and thus the inventive method can really discriminate between different species and strains, which was not possible with 16S rRNA genes sequencing.

All sequenced *E. coli* strains contained altogether 1 312 tags (including the tags to the left and to the right of the NotI recognition site) for these 3 enzymes, and among them  
30 only 139 were not unique. We can take into the account that two tags describe the same NotI site and therefore one tag can be the same but another can be different and therefore both tags still represent a unique NotI site. In such a case only 82 tags were not unique. These results demonstrate the power of the approach.

In our comparative experiments we did not use only bacterial genome sequences but the whole human genome sequences (including EST and EMBL entries). In such experiments, in the majority of the cases, NotI tags were unique even with the allowance of 1-2 sequence mismatches.

As mentioned above, the strongly advantageous feature of NotI passporting is the internal control. If a NotI site from a particular bacterial species contains for example NotI tag100 and NotI tag 101, then both tags should be obtained in approximately the same quantities. If only NotI tag100 is present, then it most probably means that NotI tag100 originates from another bacterial species.

The CODE procedure mentioned above can efficiently be applied to the NotI flanking sequences (Li et al., Proc. Natl. Acad. Sci. USA, (2002) in press). Thus, the power and sensitivity of the passporting procedure can be significantly increased by removing the most abundant species with the CODE technique (Li et al., 2001).

To be able to analyze complex microbial mixtures can be important for many applications. For instance, differences between individual composition of the normal flora will be instrumental for future analysis of how the normal flora composition is effected by diet, special foods, geographical location, colon diseases, autoimmunity, bacterial effects on colonic cancer risk, medication such as antibiotics and development of probiotics.

For this analysis we suggest to use generated restriction site tagged sequences. Hundreds of thousand tags can be produced in a short time, allowing careful analysis of thousands of bacterial species/strains (Velculesku et al., 1995). We have demonstrated that such NotI tags can be efficiently produced and that such tags have high specificity. The power of the method can be increased using the CODE subtractive procedure. We also provide a database for 'NotI passports' (as it was mentioned above it is more correct to speak about 'RSTS passports'). Such database can be used together with a NotI (RST) microarrays database (Li et al., Proc. Natl. Acad. Sci. USA, (2002) in press) as these approaches are mutually complementary. This integrated

database generates new knowledge as these two approaches are based on completely different biochemical techniques but aim to solve the same problem.

NotI-CODE subtraction

5 Prior to the present invention, the inventors developed a new genomic subtraction procedure called CODE, Cloning Of Deleted Sequences (Li et al., Biotechniques, (2001), 31: 788-793) that does not suffer from some of the limitations of RDA and RFLP subtraction. The CODE is based on the modification of the COP procedure, (Li, J., Wang, F., Zabarovska, V., Wahlestedt, C., Zabarovsky, E. R., 2000, Cloning of  
10 polymorphisms (COP): enrichment of polymorphic sequences from complex genomes. Nucleic Acids Res.), which is a new procedure for cloning single nucleotide polymorphisms. Our major objectives were to develop a simple and reproducible procedure, and to improve subtractive enrichment, thereby avoiding excessive PCR kinetic enrichment steps that often generate small DNA products.

15

In the CODE procedure, a combination of digestion with restriction enzymes, treatment with uracil-DNA glycosylase (UDG) and mung bean nuclease, PCR amplification and purification with streptavidin magnetic beads, were used to isolate deleted sequences from the genomes of two human samples. The CODE has proved to  
20 be a rather simple, efficient and robust procedure.

In the present invention two questions had to be answered:

- i) is it possible to use the CODE procedure for restriction enzymes containing CG in their recognition site and
- 25 (ii) is it possible to use NotI clones for genome wide screening for deleted, amplified and methylated NotI sites.

If the CODE procedure would work for the enzymes cutting in CpG islands, then it would be possible to clone not just deleted sequences (probably deleted by chance and  
30 without any meaning), but also genes that can be assumed as being candidate disease genes.

We suggest to use only regions surrounding NotI sites for subtraction. The novelty of

this approach is that these regions are enriched and purified using circularisation. We have designed special primers and a procedure to obtain the NotI representations (NR). The other principles for this subtraction were the same as in the CODE procedure but genomic DNA was digested with BamHI+BglII and NotI and other linkers were used to  
5 allow PCR amplification of fragments containing only NotI. Other DNA fragments were not amplified. Only two cycles of subtraction were used here.

To validate this approach, we compared a lung tumour cell line ACC-LC5 that contained a 0.7 Mb homozygously deleted region in 3p21-p22, with normal lymphocyte  
10 control DNA. We did not know if this cell line contained homozygous deletions in other chromosomes. This normal DNA is not a completely appropriate control because it was isolated from another individual. We expected cloning of polymorphic sequences as well as deleted.

An overview of the subtractive procedure is shown in Figure 1. Tester and driver DNA  
15 was digested with BamHI+BglII and self-ligated at very low concentration of DNA to form circles. Intermolecular ligation does not create any problems because the vast majority (99.99%) of these ligated molecules will be not PCR amplified in the further steps. Even rare cases, such as when these two ligated molecules contain closely located NotI sites and will be able to be PCR amplified, are useful, since they serve to  
20 normalize the representativity of different NotI surrounding sequences. Then these circles were digested with NotI. The majority (approximately 99.9%) of the circles will not be opened and thus will be omitted from further reactions. This serves also to decrease background hybridization due to illegitimate ligation of NotI linker to the DNA fragments with BamHI or BglII sticky ends.

25 The driver DNA was amplified with dUTP and unmodified primers and tester DNA were amplified with biotinylated primers in the presence of normal dNTPs. The products of DNA amplification (on average 0.5-1.5 kb) were denatured and hybridized at a ratio of 1:100 for the tester to driver DNA. After hybridization had been completed, the  
30 products were treated with UDG (which destroyed all the driver DNA) and mung bean nuclease (which digested single stranded DNA and all the non-perfect hybrids). The resulting tester homohybrids were purified, concentrated with streptavidin beads, and subjected to one more round of subtraction. The final PCR product was amplified and

cloned in the suitable vector, e.g. pBC KS(+) vector (Stratagene).

From our previous experiments we knew that the NLJ-003 and NL1-401 clones were deleted in this cell line. We isolated DNA from 10 random clones and sequenced them  
5 (to perform Southern blotting with these small inserts was impossible due to high the CG content). In this experiment scheme, only short DNA sequences (300-400 bp) were obtained, but their size can be increased using long distance PCR. Two of these clones contained NLJ-003 NotI site.

10 This experiment demonstrated that subtraction using NotI surrounding sequences is very efficient, since only 2 sites out of 10.000 NotI sites were located in the homozygously deleted region and one of them was found after analysis of only 10 clones. Other clones can be either polymorphic or/and hemizygotously deleted since when CODE procedure was applied to the same pair of driver/tester the majority of  
15 informative clones (11 of 19) fell under this category.

Thus, the present invention demonstrates that NotI-CODE procedure can be used for enzymes cutting in CpG islands.

#### 20 Use of NR for NotI clone microarrays

Thereafter we decided to check if NR after labelling with <sup>32</sup>P could be directly used for detection of deleted NotI sites. Therefore, we prepared nylon filters with immobilized DNA from NotI linking clones. These filters were hybridized to NR of ACC-LC5 (NR-A) and normal lymphocyte DNA (NR-B).

25

The results showed that these two NRs revealed different hybridization patterns: several clones hybridizing to NR-B did not hybridize to NR-A. First of all it is clear that homozygotously deleted NLJ-003 and NL1-401 were easily detected. To understand the reason why other clones failed to hybridize to NR-A, we selected 4 such clones and  
30 analysed them using Southern hybridization. Genomic DNA from ACC-LC5 and normal lymphocytes were digested either with BamHI+BglII or with BamHI+BglII+NotI, resolved by electrophoresis in agarose gel, transferred to nylon filter and hybridized to the <sup>32</sup>P labelled insert of a NotI linking clone (Figure 2:1-4). This experiment

demonstrated that all these 4 clones exhibited clear presence of a NotI recognition site in DNA from normal lymphocytes and absence of the corresponding NotI site in ACC-LC5 DNA.

5 As a next step we performed a similar experiment but used microarrays of DNA from NotI linking clones immobilized to the glass slide. The main idea of this application is shown in Figure 3. If a particular NotI site is present in the DNA then the circle will be opened with NotI and labelled. However, if this NotI site is deleted or methylated then NR will not contain the corresponding DNA sequences.

10

In a first experiment we used DNA isolated from a human-mouse microcell hybrid cell line MCH903.1 (containing the whole human chromosome 3) and MCH939.2 (chr. 3 del p14-p22). NR for MCH903.1 was labelled red and NR for MCH939.3 was labelled green. Thus sequences deleted in MCH939.2 should be red. Thereafter the deletion was  
15 precisely mapped (Figure 4A). Before the present invention, one year of work would have been needed to obtain the same results.

In a second experiment DNA from ACC-LC5 was used again to prepare NR-A and normal lymphocyte DNA was used for making NR-B. NR-A was labelled with Cy3  
20 (green) and NR-B with Cy5 (red). If both sequences are present in both NR then combined colour will be close to yellow and if some clones are deleted in ACC-LC5 then colour for these clones will be more red (Figure 4B). As it is shown in Figure 4, homozygously deleted clones NLJ-003 and NL1-401 can unambiguously be detected. Other clones showing redder colour most likely reflect the fact that in practically 100%  
25 of the cases SCLC deletion of 3p is detected. Some clones showed the same disbalance as NLJ-003 and NL1-401. This can be explained by methylation of both alleles or deletion of one allele of a NotI site and methylation (or polymorphism) of the other. Indeed, as shown in Figure 2:3-4, clones NLM-132 and NR3-077 do not contain cleavable NotI sites. In two other cases (AP20 and NRL1-1) that were also completely  
30 red, the situation is different. One allele is methylated and the other is deleted (Figure 2:5-6 and Table 5).

To further check the results of this hybridization. TaqMan probes were designed for 5



NotI linking clones. Quantitative real-time PCR was performed with these primers/probes using ABI Prism Model 7700 Sequence detector. The results of the quantitative PCR corresponded well with the NotI microarray hybridization, see Table 5 below.

5

Contamination of tumor DNA with normal DNA represents a serious problem for the identification of tumor suppressor genes. Two RCC biopsies containing 30–40% contaminating normal cells were used in a control experiment to check the sensitivity of NotI microarrays to contamination. One step of the NotI-CODE procedure was used before hybridization, and the probe was labeled with only one dye. As shown in Figure 4 (C, D), the hybridization clearly identified the two regions most frequently deleted in RCC, 3p21 telomeric (near NLJ-003) and 3p21 centromeric (near NRL1-1). Therefore, the impurity problem that can occur with tumor biopsies can be easily resolved with NotI microarrays.

10

## 15 EXAMPLES

### Cell lines and general methods

In the present invention DNA isolated from a small cell lung carcinoma cell line ACC-LC5 was used. This cell line contains homozygous 685-kb deletion in 3p21.3-p22 and was used as a source for DNA A, driver. DNA isolated from normal human lymphocytes was a control DNA (DNA B, tester).

20

Isolation of DNA, Southern transfer, hybridization, etc. were according to standard methods described in the literature. Construction of NotI linking libraries was made as described above.

25

A standard protocol was used to prepare nylon filter replicas of the gridded NotI linking clones. Nylon filters contained 100 mapped chromosome3 specific NotI linking clones and 15 random unmapped human NotI linking clones. For hybridization to nylon filter replicas of the gridded NotI clones, NR probes were <sup>32</sup>-P labeled by PCR.

30

Sequencing gels were run on ABI 310 automated sequencers (Perkin Elmer) according

to the manufacturers' protocols.

Growth of bacteria, other microbiology procedures, isolation of DNA, sequencing was performed according to standard methods.

5

The modified NotI-CODE procedure

Two oligonucleotides: NotX 5'-AAAAGAATGTCAGTGTGTCACGTATGGACGAATTCGC- 3' and NotY: 3'-AAACTTACAGTGTGTCACGTATGGCTGCTTAAGCGCCGG- 3' were used to create the NotI linker. Annealing was carried out in a final volume of 100 µl containing 20 µl of 100 µM NotX, 20 µl of 100 µM NotY, 10 µl of 10X M buffer (Boehringer Mannheim) and 50 µl of H<sub>2</sub>O. The reaction mixture was boiled for 8 min and allowed to cool slowly at room temperature (r.t.).

Two micrograms of DNA from ACC-LC5 cell line (DNA A) and normal lymphocytes (DNA B) at a DNA concentration of 50 µg/ml were digested with 20 U of *Bam*HI and 20 U of *Bgl*II (Boehringer Mannheim) at 37°C for 5 h, followed by heat-inactivation for 20 min at 65°C. Then 0.4 µg of the digested DNAs were circularized overnight with T4 DNA ligase (Boehringer Mannheim) in the appropriate buffer in 1 ml of the reaction mixture.

DNA was concentrated by precipitation in ethanol, partially filled in with for example Klenow fragment and digested with 10 U of *Not*I at 37°C for 3 h. Following digestion, *Not*I was heat inactivated and DNAs were ligated overnight in the presence of a 50 M excess of *Not*I linker at room temperature.

PCR of tester amplicon (DNA B with *Not*I linker) was performed in 100 µl of a solution containing 67 mM Tris-HCl, pH 9.1, 16.6 mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 1.0 mM MgCl<sub>2</sub>, 0.1% Tween 20, 200 µM dNTPs, 100 ng tester amplicon DNA, 400 nM of biotinylated primer NotX and 5U of Taq polymerase.

PCR of the driver amplicon (DNA A with *Not*I linker) was performed in 20 tubes using the NotX primer and the following modified conditions: dUTP (300µM) was used instead of dTTP, and 2.5mM MgCl<sub>2</sub> was used rather than 1.0mM MgCl<sub>2</sub>. The PCR

30

cycling conditions were 72°C for 5 min, followed by 25 cycles of 95°C for 1 min, 72°C for 2.5 min, and a final extension period at 72°C for 5 min. These PCR amplified tester and driver amplicons we call NotI representation (NR).

5 All PCR amplified DNA A samples were pooled (2000 µl) and mixed with 20 µl of PCR amplified DNA B (for subtraction we used a ratio of 1: 100 of DNA B to DNA A). The pooled sample was concentrated by precipitation in ethanol, purified using a JETquick PCR Purification Spin Kit (GENOMED Inc.), and dissolved in 100 µl H<sub>2</sub>O. This DNA mixture was further concentrated to 6 µl and boiled for 10 min under mineral oil.

10

Subtractive hybridization was performed for 40 h in 9 µl buffer containing 0.4 M NaCl, 100 mM Tris-HCl, pH 8.5 and 1 mM EDTA. After hybridization, the mixture was diluted to 200 µl and extracted with an equal volume of chloroform: isoamyl alcohol (24: 1) to remove the mineral oil.

15

Treatment with UDG (Boehringer Mannheim) was performed in a buffer containing 70 mM Hepes-KOH, pH 7.4, 1 mM EDTA and 1 mM dithiothreitol with 30 U UDG at 37°C for 4 hrs. Then DNA was precipitated with ethanol and dissolved in 25 µl of TE buffer. To this 3 µl of 10X MBN buffer (30 mM sodium acetate, pH 4.6, 50 mM NaCl, 1 mM zinc acetate and 0.001% Triton X-100) and 20 U of mung bean nuclease (Boehringer Mannheim) were added and incubated at 37°C for 30 min. The reaction was stopped by the addition of EDTA to a final concentration of 1 mM.

25 The subtracted DNA was purified with streptavidin coupled Dynabeads M-280 (DynaL A.S, Oslo, Norway) according to the manufacturer's instructions and dissolved in 20 µl of TE buffer. Approximately 0.5 µl of this DNA preparation was PCR amplified as described above for DNA B but using only 8 cycles, before subjecting the amplified DNA to a second round of hybridization.

30 The final subtraction product was PCR amplified, purified with JETquick PCR Purification Spin Kit (GENOMED Inc.) and digested with *NofI*. This DNA preparation was inserted into the pBC KS(+) vector (Stratagene), which was digested with *NofI* and

dephosphorylated by alkaline phosphatase (Boehringer Mannheim).

Microarray preparation, hybridization and scanning.

Microarrays were constructed essentially as described by Schena M. et al., 1996. In  
5 brief, DNA of NotI linking clones was spotted onto 3-aminopropyl-trimethoxysilane-  
coated glass microscope slides. Majority of NotI clones contained inserts 2-12 kb  
(vector part was 3.8 or 4.5 kb, see Zabarovsky et al., 1990). Qiagen-purified DNAs were  
dissolved in TE and arrayed using GMS 417 Arrayer (Genetic MicroSystems, Woburn,  
MA) with the spot density at 375 µm. The arrays were subsequently air dried,  
10 submerged in 70% EtOH for 30 min at room temperature, air dried again, and stored  
in the dark at -20°C. The microarrays described here contained 150 sequence-  
validated human chromosome 3-specific STSs in six repetitions, representing 61  
known and 49 unknown expressed sequence tags.

15 The NR probes were labelled in a PCR reaction with the NotX primer. Incorporation of  
digoxigenin or biotin was done using PCR DIG Labelling Mix (Boehringer Mannheim) or  
Biotin Reaction Mix (MICROMAX, NEN Life Science Products, Inc., Boston, MA). PCR  
products were purified using MicroSpin PCR Purification Columns (Saveen) and  
efficiency of the labelling was determined by membrane-based chemiluminescence  
20 analysis (MICROMAX, NEN).

Alternative method for preparing NR with low quality DNA was also used. According to  
this method genomic DNA was simultaneously digested with NotI and another enzyme  
or combination of enzymes not having CpG pairs in the recognition sites (e.g. Sau3A or  
25 BamHI + BglII).

After inactivation of the two enzymes, specific adaptors Sau00N and NBSgt99 were  
ligated to them:

30 Sau00N

5'-GATC CTC AAA CGC GT-3'-Amine

3'-GAG TTT GCG CAC AGC ACT GAC CCT TTT GGG ACC-5'

NBSgt99

5'-GGC CTC CAG AAA ACA TCC ACG GGC TCT AGG ATA GAT CGC-3'

3'-AG GTC TTT TGT AGG-5'

5 Thereafter, NR was prepared using PCR in the presence of Zuniv and Zgt primers. The PCR cycling conditions were 95°C for 2 min, followed by 25 cycles of 95°C for 45 sec, 65°C for 30 sec and 72°C for 1.5 min. In general, these NRs showed the same results in hybridization experiments but the background was usually higher.

10 Qualified Dig- and Bio-labelled probes were combined, denatured at 99°C, 2 min, and hybridized with denatured (0.1M NaOH, 2min, r.t.) microarrays in the Hybridization Buffer (MICROMAX, NEN) for 5h at 65°C.

The arrays were washed for 5 min at r.t. in low stringency buffer (0.06X SSC, 0.01% SDS) and developed using TSA system (MICROMAX, NEN) according to the  
15 manufacturer's protocols. In brief, we incubated microarrays with anti-DIG antibodies conjugated with horseradish peroxidase (Boehringer Mannheim) and then with Cyanine-3-Tyramide solution. After inactivation of the peroxidase in this first layer, Streptavidin-HRP Conjugate was applied and biotin residues were visualized by  
20 Cyanine-5-Tyramide.

The arrays were scanned using GMS 418 Scanner (Genetic MicroSystems, Woburn, MA), analyzed and represented by ImaGene 3.05 software (Biodiscovery). Accurate measurements of Cy3/Cy5 fluorescence ratios were obtained by taking the average of  
25 the ratios of all six spotted repetitions.

#### Quantitative real-time PCR with TaqMan probes

Oligonucleotide primers and probes were designed to amplify 5 NotI linking clones: NRL1-1 (3p21.2), NL3-001 (3p21.2 -21.32), NL1-205 (3p21.2 -21.32), NLj3 (3p21.33),  
30 924-021 (3p12.3). huBA - beta-actin gene was used as reference sequence (endogenous control). Final selection of primer and probe sequences, except huBA, was performed using the ABI Primer Express Software Version 1.5 (PE-Applied Biosystems,

Foster City, CA, USA) according to the manufacturer's instruction. TaqMan probes and primers were obtained from Perkin-Elmer. TaqMan probe consists of an oligonucleotide with a 5'- fluorescent reporter dye and a 3'-quencher dye. NLj3, NRL1-1 and hu□A probes contained FAM (6-carboxy-fluorescein), NL3-001, NL1-205 and 924-021R probes contained JOE (2,7-dimethoxy-4,5-dichloro-6-carboxy-fluorescein) as reporter dyes, located at the 5'-ends. All reporters were quenched by TAMRA (6-carboxy-N,N,N',N'-tetramethyl-rhodamine), conjugated to the 3'-terminal nucleotides. The resulting sequences are given below in Table 6

- 10 PCR reactions were carried out in 25  $\mu$ l volumes consisting of 1xPCR buffer A: 10mM Tris-HCl, 10mM EDTA, 50mM KCl, 60nM passive reference A, pH 8.3 at room temperature; 3.5mM MgCl<sub>2</sub>, 200 $\mu$ M dATP, dGTP, dCTP, 400  $\mu$ M dUTP, 100nM TaqMan probe, forward and reverse primers in appropriate concentrations, 0.025 unit/ $\mu$ l AmpliTaq Gold DNA polymerase, 0.01 unit/ $\mu$ l AmpErase and 5  $\mu$ l of
- 15 appropriate diluted DNA template. H<sub>2</sub>O was added to 25  $\mu$ l of total volume. PCR were performed using ABI Prism® Model 7700 Sequence Detector. The reactions were done in triplicate for each sample in the same or separate tubes.

- The primer limitation experiments were performed for multiplex PCR with more than
- 20 one primer pair in the same tube (ABI PRISM 7700 Sequence Detection System. User Bulletin no.2. Relative quantitation of Gene Expression. PE Applied Biosystems, 1997). Thermal cycling conditions consisted of 2 min at 50°C, 10 min at 95°C, followed by 40 cycles of 15 s at 95°C and 1 min at 60°C.

- 25 Cycle threshold (C<sub>T</sub>) determinations (i.e. calculations of the number of cycles required for reporter dye fluorescence resulting from the synthesis of PCR products to become significantly higher than background fluorescence levels) were automatically performed by the instrument for each reaction.

- 30 Details concerning the theory and derivation of the comparative C<sub>T</sub> method ( $\Delta\Delta C_T$  method) for target sequence quantitative assessment has been published (ABI PRISM 7700 Sequence Detection System. User Bulletin no.2. Relative quantitation of Gene Expression. PE Applied Biosystems, 1997). This method is dependent upon the inverse

exponential relationship that exists between starting quantity (number) of target sequence copies in the reactions and corresponding CT determinations by the ABI7700 system: the more copies, the less value CT (ABI PRISM 7700 Sequence Detection System. User Bulletin no.2. Relative quantitation of Gene Expression. PE Applied Biosystems, 1997). We used an approach referred to as the comparative cycle threshold (CT) method to determine target sequence quantity of tumour sample – ACC-LC5, (target) relative to those in the sample for comparison – normal DNA, (calibrator) and compared with an endogenous control sequence - beta-actin (reference) in both samples. For amplicons designed and optimized according to PE Applied Biosystems guidelines, efficiency is close to 100 %. In this case, the amount of target (copy number), normalized to an endogenous reference and relative calibrator, is given by:  $N_{ACC-LC5} / N_{calibrator} = 2^{-\Delta\Delta C_T}$ . The calculation  $\Delta\Delta C_T$  involves subtraction of mean reference sequence  $C_T$  values from mean target sequence  $C_T$  for ACC-LC5 and CBMI, to obtain values  $\Delta C_{T^{ACC-LC5}} = C_{T^{target}} - C_{T^{actin}}$  and  $\Delta C_{T^{norm}} = C_{T^{target}} - C_{T^{actin}}$ . The values  $\Delta C_{T^{norm}}$  are then subtracted from values  $\Delta C_{T^{ACC-LC5}}$  to obtain  $\Delta\Delta C_T$ . The range given for all probes relative to  $\beta$ -actin was determined the expression:  $2^{-\Delta\Delta C_T}$  with  $\Delta\Delta C_T + s$  and  $\Delta\Delta C_T - s$ , where  $s$  = the standard deviation of the  $\Delta\Delta C_T$  value.

For the  $\Delta\Delta C_T$  calculation to be valid, the efficiency of the target amplification and efficiency of the reference amplification must be approximately equal. Before using the  $\Delta\Delta C_T$  method for quantitative assessment a validation experiment was performed (ABI PRISM 7700 Sequence Detection System. User Bulletin no.2. Relative quantitation of Gene Expression. PE Applied Biosystems, 1997). The performed validation experiments demonstrated that efficiencies of these targets and references are approximately equal for chosen dilutions. In this case we can use the  $\Delta\Delta C_T$  calculations for the relative quantitation of target without using standard curves.

Data analysis was done using Sequence Detection System (SDS) software (PE-Biosystems).

#### The NotI-passporting procedure

Two oligonucleotides, BfocII: 5'-ggatgaaaactgga-3' and Z98NOT: 3'-gtcgtgactgggaaaaccctggcctacttttgacctccgg-5' were used to create the NotI linker.

Two micrograms of bacterial DNA at a concentration of 50 µg/ml were digested with 20 U NotI (Roche Molecular Biochemicals) at 37 °C for 2 h and heat-inactivated for 20 min at 85 °C. Then, 0.4 µg of the digested DNA was ligated to NotI linker (50 M excess) overnight with T4 DNA ligase (Roche Molecular Biochemicals) in the appropriate buffer in 100-µl reaction mixtures. The DNA was then concentrated by precipitation in ethanol and digested with 10 U BpmI at 37°C for 3 h.

Following digestion, BpmI was heat-inactivated and the DNA was ligated overnight in the presence of a 50 M excess of the ZNBpm linker at room temperature. Two nucleotides, the Zamine: 5'-ctcaaaccgt-3' and the Z2\_univer: 3'-Nngagtttggcacagcactgacccttttgggacc-5' were used to create the ZNBpm linker.

The sample was then purified using a JETquick PCR Purification Spin Kit (GENOMED Inc.), and dissolved in 100 µl TE. One microliter of this sample was PCR amplified with Z1 univer (3'-gagtttggcacagcactgacccttttgggacc-5') and antiuniver (5'-cagcactgacccttttgggacc-3') primers.

PCR was performed in 40 µl solution containing 67 mM Tris-HCl (pH 9.1), 16.6 mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 2.0 mM MgCl<sub>2</sub>, 0.1% Tween 20, 200 µM dNTPs, 3 µl PCR pool, 400 nM of each primer, and 5 U Taq DNA polymerase. The PCR cycling conditions were 95 °C for 1.5 min, followed by 25 cycles of 95 °C for 1 min, 60 °C for 1 min, with 72 °C for 0.5 min, with a final extension period at 72 °C for 3 min.

The final product was purified with the JETquick PCR Purification Spin Kit (Genomed GmbH) and cloned using TOPO TA Cloning kit (Invitrogen AB, Sweden). Sequencing gels were run on ABI 377 automated sequencers (Perkin Elmer), according to the manufacturers' protocols, using standard primers.

For the analysis of the complex flora composition, we suggest using only some specific fragments of the genomes (e.g. NotI representations, NotI tags, NotI linking clones, etc.). Thus we do not aim to sequence all genomes or study all genes. We append



special signatures for the particular microorganism/genes and analyze these signatures in different samples of colon flora. In the present invention study work we have analyzed the use of short sequence tags appended to NotI or other restriction enzyme recognition site. The collection of NotI tags represents NotI sequence passport  
5 or in short NotI passport and NotI passporting means creation of NotI tags/passports. The naming is based on the initially used enzymes, but the methods can be adapted to other restriction enzymes as well.

The general design of the experiment is as follows (Figure 5). DNA generated from  
10 faecal samples and surgical specimens are digested with NotI and ligated to special linker containing BpmI recognition site. Then DNA is digested with BpmI, ligated to the special linkers and PCR amplified. We have proved that in these conditions only specific 85 bp NotI-BpmI fragments are amplified (Figure 6). After digestion with BpmI and FokI this fragment will generate 24 bp fragments which represent particular NotI  
15 sites. From here it is possible to work in two directions.

a) concatemer strategy

The 24 bp units will be ligated into the concatemers of about 1.000 bp size, cloned and sequenced. Each sequencing reaction will give information about 20 - 50 NotI sites.

b) oligomer strategy.

20

New high-throughput sequencing techniques, such as pyrosequencing or massively parallel signature sequencing have been developed recently. They allow one person to produce many thousands sequences per day. However, these sequences are very short 20-40 bp and suit our needs well, whereby NotI passport for the particular specimen  
25 can be produced. Comparing these passports from e.g. different individuals or from the same individual before and after drug treatment we find the difference between them. This information in some cases can be directly used to make conclusions. In other cases, using these sequences we can identify NotI linking clones which are different between two samples. These clones can be used for further analysis, e.g. finding the  
30 genes which are responsible for a certain medical condition (e.g. cancer, aging etc.) or sequencing/isolation of the required microorganism.

Table 1. Comparison of different microarrays to study genome copy number changes and methylation.\*

Method/ Feature	cDNA	CGH (BACs, P1, PACs)	Representation	SNP	CGI	RST (NotI microarrays)
Homozygous deletions	Low	Yes	Yes/NO	NO	Yes/NO	Yes
Hemizygous deletions	Low	Yes	NO	NO	NO	Yes
LOH	NO	NO	Yes	Yes	Yes/NO	Yes
Amplifi- cation	Low/ Medium	Yes	Yes	NO	Yes	Yes
Methylation	NO	NO	NO	NO	NO	Yes
Number of available markers	More than 40.000	10.000- 30.000	1.500 (polymorphic BglII fragments per genome)	1.300 (can be increased)	1.500 (polymorphic BglII fragments per genome)	10.000- 20.000
Connection to genes	Direct	Indirect	NO (indirect)	NO (indirect)	NO (indirect)	Direct
Main disadvan- tages	Low sensitivity and precision; small size of the insert; high background, several EST markers should be used to determine copy number; 100% DNA is labeled (many repeats)	Very expensive, difficulties to work with large-insert vectors: low yield, rearrange- ments; 100% DNA is labeled (many repeats)	Not convenient for large-scale screening; small size of the inserts; only 1% polymorphic sites; unknown location; 2.5% of all DNA is labeled, unknown purification from repeats.	High sensitivity to normal cell contamination, short hybridizing sequences; many reactions and primers are needed; expensive; less than 30% of markers are polymorphic	Not convenient for large-scale screening; small size of the inserts; only 1% polymorphic sites; unknown location; 2.5% of all DNA is labeled, unknown purification from repeats.	High CG content; small size of the inserts; discrimination between LOH, deletions/amplifi- cations and methylation

Main advantages	Direct connection to RNA profiling	Good to check copy number changes	Can be easily adopted for small scale experiments (small genomes)	Very small fraction of the genome is labeled; good to check LOH	Can be easily adopted for small scale experiments (small genomes)	Methylation detection; up to 45% of clones are polymorphic, easy to solve normal cell contamination, one reaction and one pair of primers are used; comparatively cheap, good to check LOH and copy number changes, only 0.1-0.2% of the genome is labeled; 10 fold purification from repeats; direct connection to genes; simultaneous detection of different aberrations associated with cancer development
-----------------	------------------------------------	-----------------------------------	---	---	---	---

\* Efficiency of the method to detect the particular feature

Table 2 Comparison of NotI and CGI microarrays

<b>Feature</b>	<b>NotI microarrays</b>	<b>CGI-microarrays</b>
Uncomplete restriction digestion	No effect	Artificial result
Specificity of labeling	0.1-0.5% of the total human DNA	100% total human DNA
Repeats	10% compared to the average in human genome	Approximately the same as in average
rRNA genes	No	Yes
Homozygous deletions	Yes	No
Hemizygous deletions	Yes	No
Hemizygous methylation	Yes	No
Oligo microarrays	Yes	???
Homozygous methylation in cancer cells	Yes	Yes
Quality of clones	All sequenced, all contain genes	Partly sequenced, many repeated sequences and repeats like LINE etc.
Number of available clones	> 5.000	Unknown

5

10

15

20

Table 3. The number of recognition sites for rare cutting restriction enzymes in selected bacterial genomes

	GENOME	SIZE	NotI	PacI	PmeI	SbfI	SgfI	SgrAI	Sse2321	SwaI
1	<i>Bacillus subtilis</i>	4,2	81	89	89	51	51	157	52	176
2	<i>Borrelia burgdorferi</i>	1,5	1	234	37	8	0	2	0	548
3	<i>Campylobacter jejuni</i>	1,6	0	91	42	13	5	1	0	526
4	<i>Chlamydia pneumoniae</i> AR39	1,2	2	59	10	21	13	4	1	60
5	<i>Deinococcus radiodurans</i> R1	3,3	15	1	4	28	7	645	164	1
6	<i>Escherichia coli</i> K12	4,6	23	143	87	68	222	548	31	117
7	<i>Escherichia coli</i> O157:H7	5,5	36	165	92	108	239	642	34	126
8	<i>Helicobacter pylori</i> 26695	1,7	7	32	35	4	88	61	12	67
9	<i>Helicobacter pylori</i> J99	1,6	14	34	43	4	87	66	15	76
10	<i>Lactococcus lactis</i> subsp. <i>lactis</i>	2,4	3	176	47	17	2	11	0	235
11	<i>Rickettsia prowazekii</i>	1,1	1	239	20	10	1	4	0	229
12	<i>Staphylococcus aureus</i> Mu50	2,9	0	440	83	12	5	12	2	602
13	<i>Streptococcus pneumoniae</i> R6	2,0	1	40	25	30	1	9	0	51
14	<i>Synechocystis</i> PCC6803	3,6	44	192	104	40	3160	182	18	167
15	<i>Vibrio cholerae</i>	4,0	73	103	117	37	203	199	24	104

Table 4. Specificity of restriction tags in *E. coli* and *H. pylori* strains.

Species	Strain	Cutting sites in the genome					Unique for the species					Unique for the strain				
		Pme I	Sbfl	NotI	Tota I		Pme I	Sbfl	NotI	Tota I		Pme I	Sbfl	NotI	Tota I	
<i>Escherichia coli</i>	K12 (4.6 Mb)	87	68	23	178		74	61	20	155		25	26	6	57	
	O157H7 (5.5 Mb)	92	108	36	236		77	90	34	201		28	55	20	103	
<i>Helicobacter pylori</i>	26695 (1.7 Mb)	35	4	7	46		35	4	7	46		21	2	4	27	
	J99 (1.6 Mb)	43	4	14	61		43	4	14	61		33	2	11	46	

Table 5. Relative quantitative measurements using comparative ( $\Delta\Delta C_T$ ) method for normal lymphocyte DNA and ACC-LC5 cell line

Target/colour	Location	$N_{ACC-LC5} / N_{norm} = 2^{-\Delta\Delta C_T}$	Comments
924-021/yellow	3p12.3	0.94 (0.83 – 1.05)	No changes
NRL1-1/red	3p21.2	0.51(0.41-0.62)	Initial target sequence copy number in ACC-LC5 is half of what is obtained in CBMI (hemizygous deletion)
NL3-001/yellow	3p21.2 - 21.32	1.12 (0.98 – 1.26)	No changes
NL1-205/yellow	3p21.2- 21.32	1.25 (0.75 – 1.74)	No changes
NLj3/red	3p21.33	0.00	Zero sequence copy number (homozygous deletion)

Table 6. TaqMan probe, primer sequences and product lengths

Target	Oligonucleotide	Sequence (5' → 3')	Amplicon, bp
924-021 (3p12.3)	924-021, probe primer(F) primer(R)	TGCTGGCCACAGGCCCTGC TGCATGTGCCAGTGTGATAAA GTGTTGTGAGCCCTGGGAA	52
NRL1-1 (3p21.2)	NRL1-1, probe primer(F) primer(R)	AGCCTGAGCTGGGCAGACAGTTTCC CAGCCCCACGGTCACTTC GCCAAAACAGACCCAGCCT	74
NL3-001 (3p21.2 -21.32)	NL3-001, probe primer(F) primer(R)	CCCCAGAAACGCGCGGGC CTTGCCATCTGCAATTCCT CTCCATGAGGCTGTGGGAAG	60
NL1-205 (3p21.2 -21.32)	NL1-205, probe primer(F) primer(R)	GCGGCTGGCTCTGCGC ATGAGGCTCTTTCCCATGCC GCCGGATTCAGGATGCTTT	63
NLj3 (3p21.33)	NLj3, probe primer(F) primer(R)	CTGGCGGAGAGACTGGGAGCGA CAGAGTGCGTGTGCCGACT ACAACCTCTCTGCGGGCGT	125
hu?A (control) 7 chromosome	hu?A, probe primer(F) primer(R)	ATGCCCCCCCCCATGCCATCCTGCGT TCACCCACACTGTGCCCATCTACGA CAGCGGAACCGCTCATTGCCAATGG	295



## REFERENCES

- Bicknell, D. C., Markie, D., Spurr, N. K., Bodmer, W. F., 1991. The human chromosome content in human x rodent somatic cell hybrids analyzed by a screening technique using Alu PCR. *Genomics* 10, 186-192.
- 5 -Brookes, A. J., 1999. The essence of SNPs. *Gene* 234, 177-186.
- Costello JF, Fruhwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, Wright FA, Feramisco JD, Peltomaki P, Lang JC, Schuller DE, Yu L, Bloomfield CD, Caligiuri MA, Yates A, Nishikawa R, Su Huang H, Petrelli NJ, Zhang X, O'Dorisio MS, Held WA, Cavenee WK, Plass C. Aberrant CpG-island methylation has non-random and tumour-
- 10 type-specific patterns. *Nat Genet* 2000 24(2): 132-138
- Espinosa-Urgel, M., Kolter, R., 1998. *Escherichia coli* genes expressed preferentially in an aquatic environment. *Mol. Microbiol.* 28, 325-332.
- Ishikawa, S., Kai, M., Tamari, M., Takei, Y., Takeuchi, K., Bandou, H., Yamane, Y., Ogawa, M., Nakamura, Y., 1997. Sequence analysis of a 685-kb genomic region on
- 15 chromosome 3p22-p21.3 that is homozygously deleted in a lung carcinoma cell line. *DNA Res.* 4, 35-43.
- Kaiser, C., Von Stein, O., Laux, G., Hoffmann M., 1999. Functional genomics in cancer research: identification of target genes of the Epstein-Barr virus nuclear antigen 2 by subtractive cDNA cloning and high-throughput differential screening using high-
- 20 density agarose gels. *Electrophoresis* 20, 261-268.
- Li, J., Wang, F., Zabarovska, V., Wahlestedt, C., Zabarovsky, E. R., 1999. Cloning of polymorphisms (COP): enrichment of polymorphic sequences from complex genomes. *Nucleic Acids Res.*, in press.
- Lisitsyn, N., Lisitsyn, N., Wigler, M., 1993. Cloning the differences between two
- 25 complex genomes. *Science* 259, 946-951.
- Lisitsyn, N.A., Segre, J.A., Kusumi, K., Lisitsyn, N.M., Nadeau, J.H., Frankel, W.N., Wigler, M.H., Lander, E.S., 1994. Direct isolation of polymorphic markers linked to a trait by genetically directed representational difference analysis. *Nat. Genet.* 6: 57-63.
- Lisitsyn, N.A., Lisitsina, N.M., Dalbagni, G., Barker, P., Sanchez, C.A., Gnarr, J.,
- 30 Linehan, W.M., Reid, B.J., Wigler, M.H., 1995. Comparative genomic analysis of tumours: detection of DNA losses and amplification. *Proc. Natl. Acad. Sci. U S A* 92: 151-155.

- Parikh, V. S., Morgan, M. M., Scott, R., Clements, L. S., Butow, R. A., 1987. The mitochondrial genotype can influence nuclear gene expression in yeast. *Science* 235: 576-580.
- Rosenberg, M., Przybylska, M., Straus, D., 1994. 'RFLP subtraction': a method for making libraries of polymorphic markers. *Proc. Natl. Acad. Sci. U S A*, 91: 6113-6117.
- 5 -Sambrook, J., Fritsch, E. F., Maniatis, T., 1989. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Sugai, M., Kondo, S., Shimizu, A., Honjo, T., 1998. Isolation of differentially expressed genes upon immunoglobulin class switching by a subtractive hybridization method using uracil DNA glycosylase. *Nucleic Acids Res.* 26, 911-918.
- 10 -Sugimura T, Ushijima T. Genetic and epigenetic alterations in carcinogenesis. *Mutat Res* (2000) 462(2-3): 235-246.
- Yamakawa, K., Takahashi, T., Horio, Y., Murata, Y., Takahashi, E., Hibi, K., Yokoyama, S., Ueda, R., Takahashi, T., Nakamura, Y., 1993. Frequent homozygous deletions in lung cancer cell lines detected by a DNA marker located at 3p21.3-p22. *Oncogene* 8, 327-330.
- 15 -Yan PS, Chen CM, Shi H, Rahmatpanah F, Wei SH, Caldwell CW, Huang TH. Dissecting complex epigenetic alterations in breast cancer using CpG island Microarrays. *Cancer Res* 2001 61(23): 8375-8380
- 20 -Zabarovsky, E. R., Boldog, F., Thompson, T., Scanlon, D., Winberg, G., Marcsek, Z., Erlandsson, R., Stanbridge, E. J., Klein, G., Sumegi, J., 1990. Construction of a human chromosome 3 specific NotI linking library using a novel cloning procedure. *Nucleic Acids Res.* 18, 6319-6324.
- Zabarovska, V., Li, J., Muravenko, O., Fedorova, L., Ernberg, I., Wahlestedt, C., Klein, G. and Zabarovsky, E. R. CIS - cloning of identical sequences between two complex genomes. *Chromosome Research*, in press.
- 25 -Allikmets et al.: NotI linking clones as tools to join physical and genetic mapping of the human genome. *Genomics*, (1994) 19: 303-309.
- Kashuba et al.: Analysis of NotI linking clones isolated from human chromosome 3 specific libraries. *Gene*, 1999, 239: 259-271.
- 30 -Katouli et al. Composition and diversity of intestinal coliform flora influence bacterial translocation in rats after hemorrhagic stress. *Infection and immunity* 62: 4768-4774,

1994

-Klein, J. (1999). *Batrachomyomachia*: frog 1, mice 0. *Scand J Immunol*, 49: 11-13.

-Li J. et al.: COP - a new procedure for cloning single nucleotide polymorphisms.

*Nucleic Acids Res.* (2000) 28, e1,p. i-v.

5 -Li et al.: CODE: a new genomic subtraction method for cloning deleted sequences.

*Biotechniques*, (2001), 31: 788-793.

-Midtvedt, T. (1999). In L. Å. Hansson and R. H. Yolken (eds.), *Microbial functional Activities*. Lippincott-Raven, Philadelphia, Vol. Nestle Nutritional Workshop Series, pp. 79-96.

10 -Ronaghi et al. Pyrosequencing – a DNA sequencing method based on real-time pyrophosphate detection. *Science* (1998) 281: 363-365.

-Sandberg et al.: Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Res.* 2001, 11:1404-1409.

Velculesku et al.: Serial analysis of gene expression. *Science* (1995) 270:484-487.

15 -Zabarovska et al.: Slalom libraries: a new approach to genome mapping and sequencing. *Nucleic Acids Res.* (2002) 30 (e6): 1-8.

-Zabarovsky et al.: Construction of a human chromosome 3 specific NotI linking library using a novel cloning procedure. *Nucl. Acids Res.*, 1990, 18:6319.

20 -Zabarovsky et al.: A new strategy for mapping the human genome based on a novel procedure for constructing jumping libraries. *Genomics*, 1991, 11: 1030-1039.

-Zabarovsky et al.: NotI clones in the analysis of human genome. *Nucl. Acids Res.*, (2000) 28: 1635-1639.

-Zabarovsky et al.: Novel techniques to identify the species composition of complex microbial systems: restriction site tagged microarrays (RST) and NotI signatures.

25 *Microecology and Therapy*, in press.

**CLAIMS**

1. Method for preparing nucleic acid or and/or modified nucleic acid reference material bound to a solid phase, comprising the steps of:
    - 5 -digesting nucleic acid and/or modified nucleic acid reference material using biochemical and/or chemical approaches, to obtain sequence fragments surrounding a specific recognition site,
    - selecting said nucleic acid and/or modified nucleic acid sequence fragments associated with a specific recognition site.
  - 10 2. Method according to claim 1, wherein said reference material is digested by a first restriction enzyme and/or one or more second restriction enzymes.
  3. Method according to claim 2, wherein the restriction enzymes are endonucleases.
  - 15 4. Method according to claim 3, wherein the recognition sites of the first endonuclease is scarcely distributed along said genomic material and is located adjacent to gene sequences, and the recognition sites of said one or more second restriction endonucleases are more frequently occurring along said genomic material than the sites of the first endonuclease.
  - 20 5. Method of claim 4, wherein the digestion by the first and second restriction endonucleases are performed simultaneously, and different linkers are ligated to the ends resulting from cutting by the first and second restriction endonucleases, respectively, which linkers are designed such that when primers are added in order to make PCR reactions, only the fragments containing ends resulting from cutting by the first restriction endonuclease will be amplified.
  - 25 6. Method of claim 4, wherein the reference material is first digested by the one or more second restriction endonucleases, the ends of the thus obtained fragments are self-ligated into the form of circular nucleic acid and/or modified nucleic acid molecules, and any linear fragments remaining after self-ligation are
  - 30
-

inactivated before digestion with the first restriction endonuclease, whereby the linear fragments resulting from the digestion by the first endonuclease are subjected to PCR amplification.

- 5 7.Method of any claims 2-6, wherein the first restriction endonuclease is NotI, or any other restriction endonuclease, the restriction sites of which occurs in proximity to CpG islands in the genomic material.
- 10 8.Method of of any claims 2-6, wherein the first restriction endonuclease is NotI, PmeI or SbfI, or a combination of two or more of said endonucleases, and the second endonuclease is BamHI, BclI, BglII or Sau3A, or a combination of two or more of said endonucleases.
- 15 9.Method according to any of the preceding claims, wherein said nucleic acid and/or modified nucleic acid reference material is selected from RNA, DNA, peptides or modified oligonucleotides, or a combination of two or more of said materials.
- 20 10.Method according to any of the preceding claims, wherein the solid phase is a glass slide, coded beads, cellulose, such as nitrocellulose, or filters.
- 25 11.Method of any of the previous claims, wherein the genomic material is derived from one or more humans, from different locations in the body/bodies and at the same or different points in time.
- 30 12.Method of any of the previous claims, wherein the genomic material is derived from bacteria from the gut, skin or other parts of the human body.
- 13.Method of any of the previous claims, wherein the genomic material is derived from any organism, bacteria, animal, or plant, or product produced therefrom, or from any substance wherein genomic material can be contained, especially air and water.

14. Fragments obtained by means of the method of any of the preceding claims.

5 15. Nucleic acid or and/or modified nucleic acid microarray containing the fragments of claim 14.

10 16. Representation of the genome, or of a part thereof, of an organism, comprising multiple copies of the nucleic acid and/or modified nucleic acid fragments, or a selection thereof, obtained by means of the method of any of the claims 1 - 13.

17. Representation of claim 16, wherein the representation in liquid form is hybridized to the nucleic acid and/or modified nucleic acid fragments present in the form of said solid phases.

15 18. Use of the representation of claim 16 or 17 in discriminating between different genomes, detecting methylations, deletions, mutations and other changes within genomic material obtained from the same individual at different points of time, or in the genomic material obtained from one individual as compared to a standard representation obtained from at least one other individual, or a  
20 combination thereof.

19. Use of representation of claim 16 or 17 for:

- 25 -studying methylation and copy number changes in eukaryotic genomes for diagnosis, prognosis, identification of cancer causing genes, etc,
- genotyping different microorganisms (viruses, prokaryotic, eukaryotic),
- studying biocomplexity and diversity of complex biological systems, i.e. human gut, bacterial flora in water, food, air resources,
- identifying pathogenic organisms in different sources including complex biological mixtures,
- 30 -producing passports (images of microarrays hybridizations, databases containing tag sequences) for different purposes: to describe organisms at different conditions, i.e. different ages, disease/healthy,

infected/uninfected etc,

-identifying new organisms, e.g. bacterial species,

-producing microarrays (DNA- and oligo-based) to study all above  
described features,

5 -verification and maintenance of large biological collection/banks, i.e.  
verifying cell lines and individual organisms for higher organisms and  
confirming the purity of the particular strain for microbial species,

-producing kits for labeling and hybridization with microarrays,

-producing kits for making sequence tagging (passporting), and

10 -producing oligo microarrays to analyze sequence tags.

20. NotI CODE genomic subtraction method based on the use of fragments of claim 14.

Fig. 1

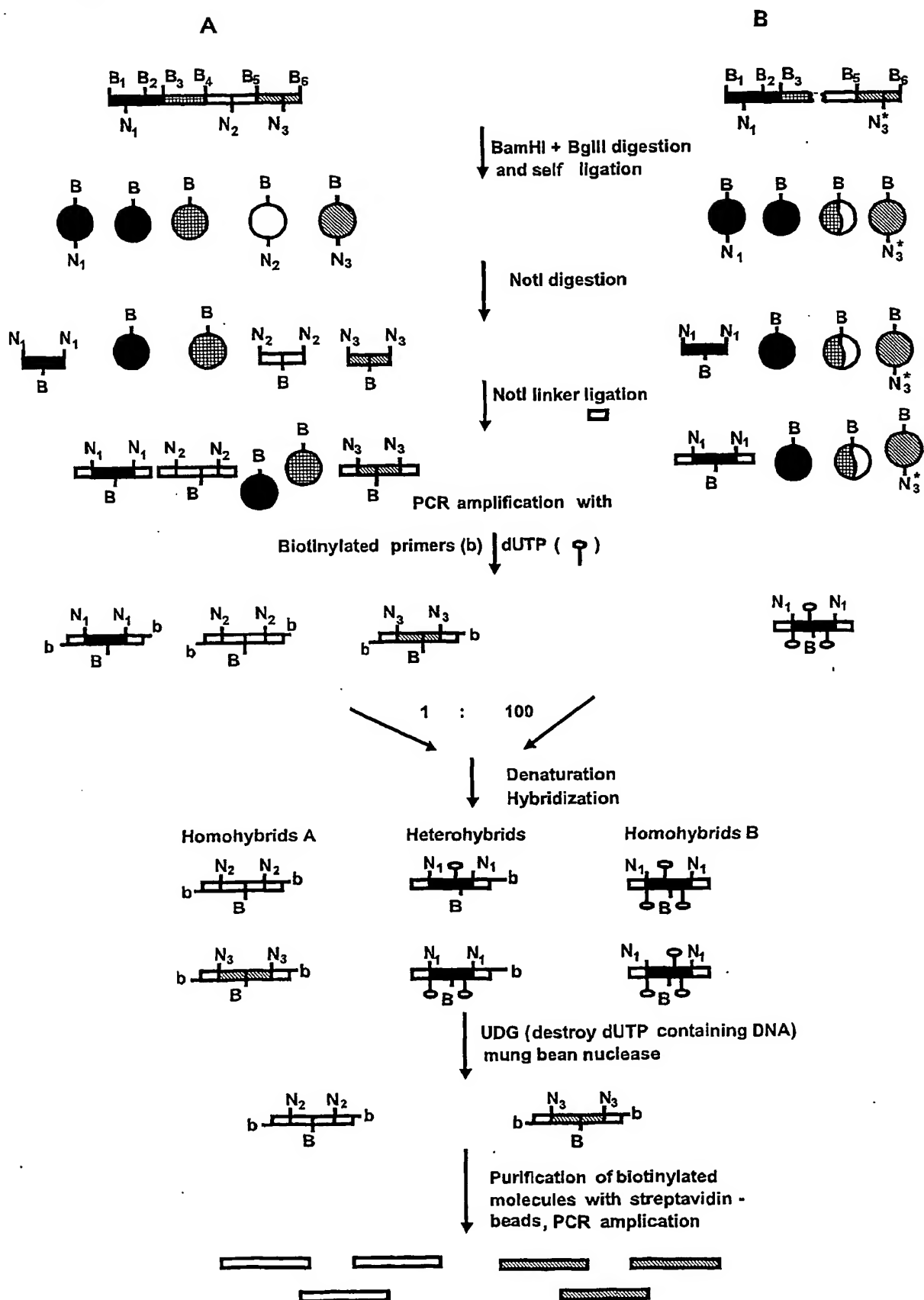




Fig. 2

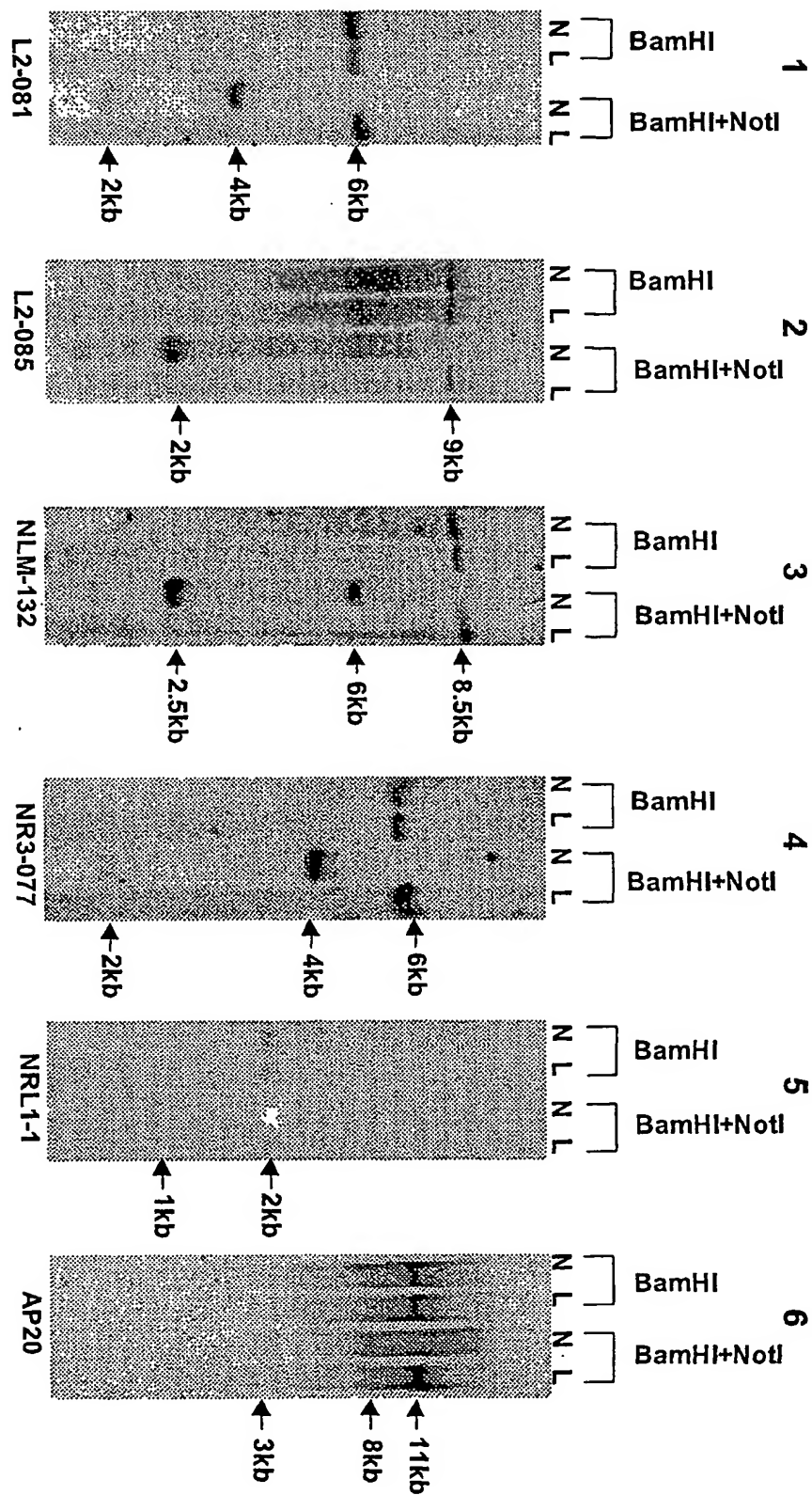


Fig. 3

# Figure 3. NotI representation protocol

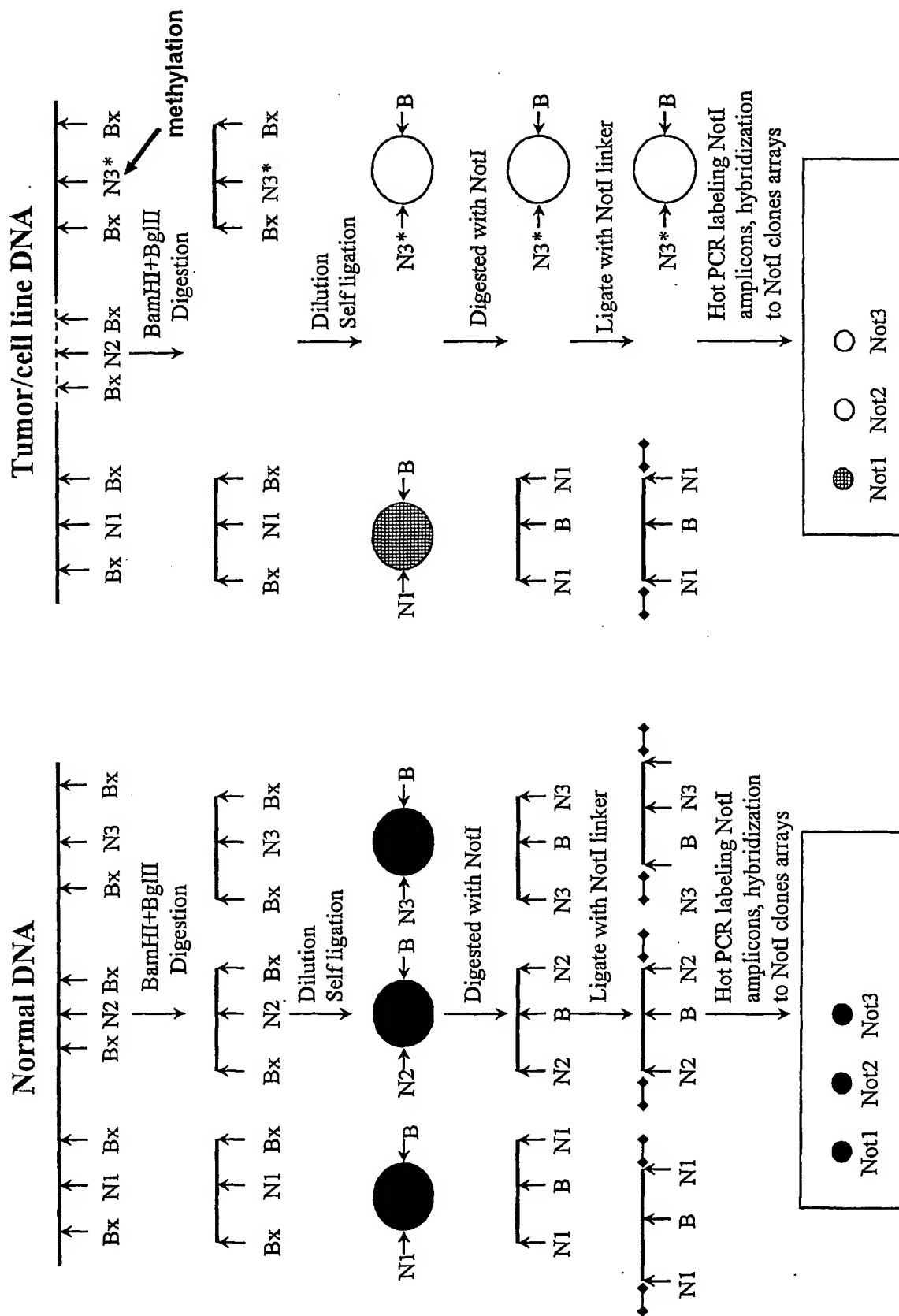


Fig. 4

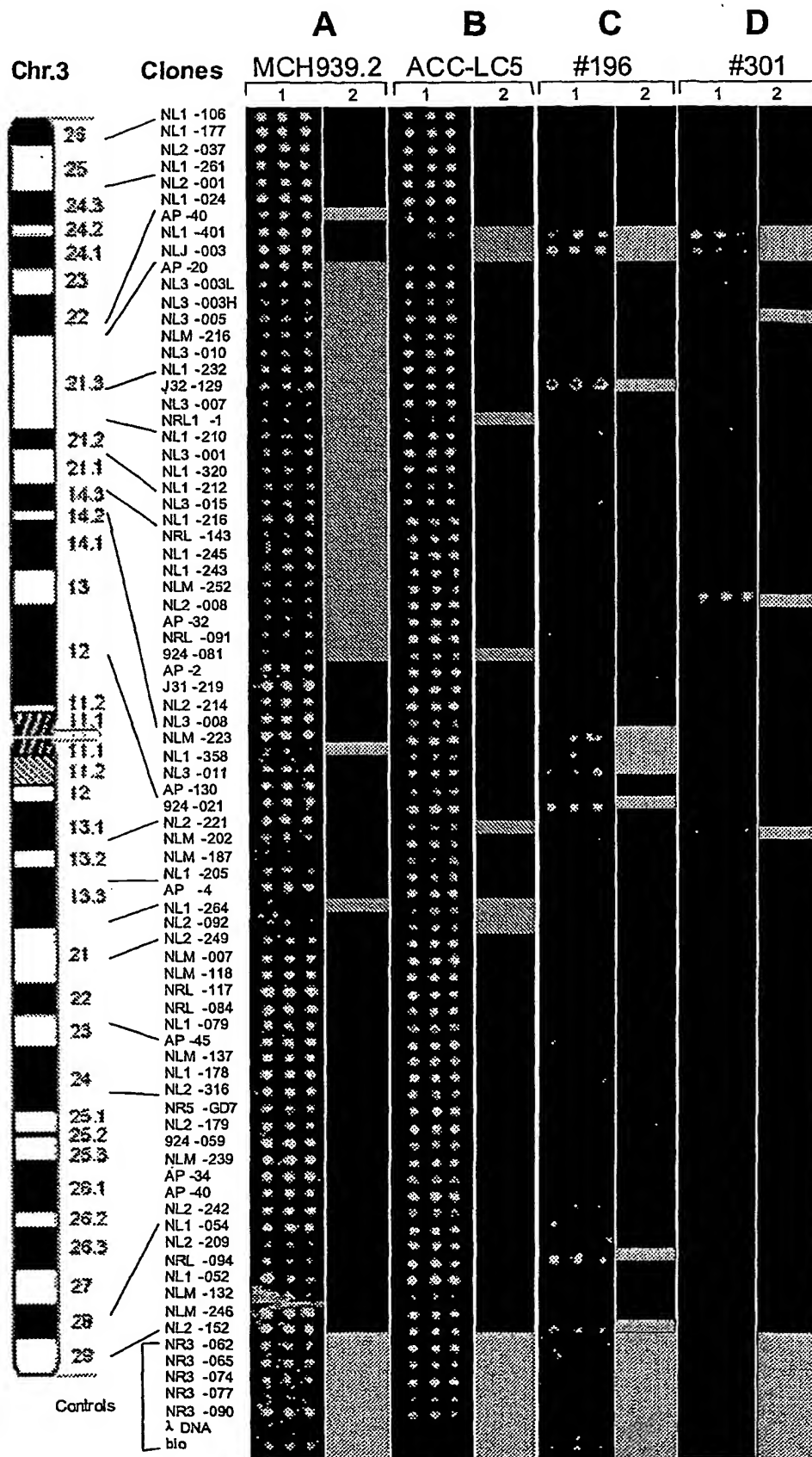
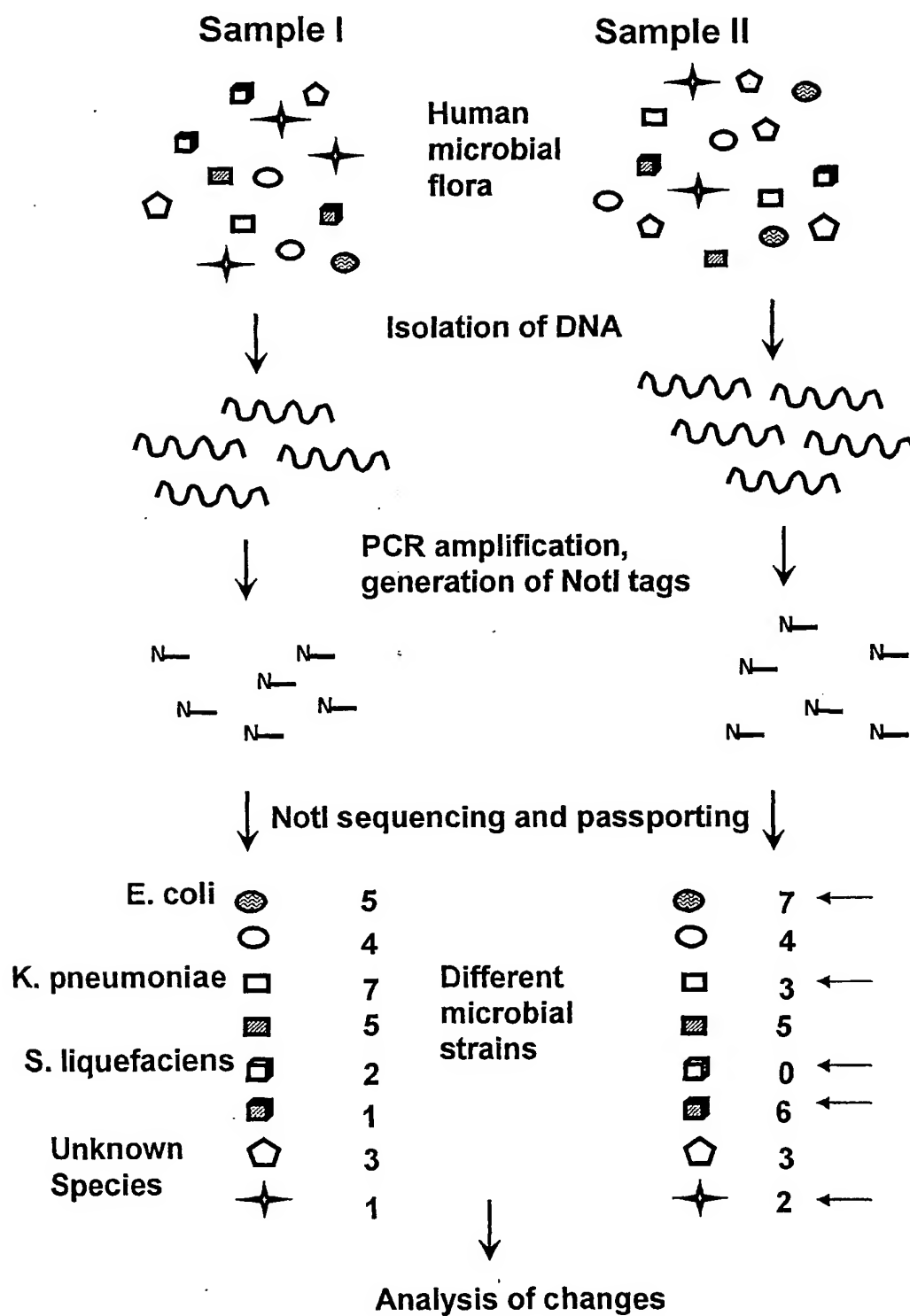


Fig. 5





## INTERNATIONAL SEARCH REPORT

International application No.

PCT/SE 02/00788

## A. CLASSIFICATION OF SUBJECT MATTER

IPC7: C12Q 1/68

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC7: C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-INTERNAL, PAJ, WPI DATA

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Genome Research, Volume 10, 2000, Robert Lucito et al: "Detecting Gene Copy Number Fluctuations in Tumor Cells by Microarray Analysis of Genomic Representations", pages 1726-1736, page 1727, left column, paragraph 2; page 1734, left column, paragraph 2, page 1735, left column, paragraph 4	1-5,9-19
Y	--	6-8,20

☒ Further documents are listed in the continuation of Box C.☒ See patent family annex.

\* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&amp;" document member of the same patent family

Date of the actual completion of the international search

Date of mailing of the international search report

21 August 2002

04-09-2002

Name and mailing address of the ISA/

Swedish Patent Office

Box 5055, S-102 42 STOCKHOLM

Facsimile No. +46 8 666 02 86

Authorized officer

SARA NILSSON/BS

Telephone No. +46 8 782 25 00

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/SE 02/00788

## C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 9923256 A1 (COLD SPRING HARBOR LABORATORY), 14 May 1999 (14.05.99), page 12, line 19 - line 27; page 7, line 10 - page 8, line 31; page 19, line 8 - line 34, page 21, line 17 - line 21; page 22, line 3 - line 17; page 29, line 4 - line 10	1-5,9-19
Y	--	6-8,20
Y	Genomics, Volume 16, 1993, Eugene R. Zabarovsky et al: "Alu-PCR Approach to Isolating Not-Linking Clones from the 3p14-p21 Region Frequently Deleted in Renal Cell Carcinoma", pages 713-719, page 714, right column, paragraph 2; page 717, right column, paragraph 2; page 718 last sentence	6-8
Y	--	
Y	Nucleic Acids Research, Volume 28, No. 7, 2000, Eugene R. Zabarovksy et al: "Not clones in the analysis of the human genome", pages 1635-1639, page 1635, left column, paragraph 2; page 1635, right column, paragraph 3; page 1636, left column, paragraph 4	7-8
Y	--	
Y	WO 9842871 A1 (BOEHRINGER MANNHEIM CORPORATION), 1 October 1998 (01.10.98), page 29, line 30 - page 30, line 22; page 28, line 5 - line 7, page 52, claim 47, figure 1; page 32, line 1 - line 5	20
A	--	
A	Cancer Detection and Prevention, Volume 20, No.1, 1996, Eugene R. Zabarovsky et al: "Not Jumping and Linking Clones as a Tool for Genome Mapping and Analysis of Chromosome Rearrangements in Different Tumors", pages 1-10, see especially page 4	1-20
	--	

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/SE 02/00788

## C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Carcinogenesis, Volume 21, No. 3, 2000, Joe W. Gray et al: "Genome changes and gene expression in human solid tumors", pages 443-452  --	1-20
A	International Journal of Oncology, Volume 17, 2000, Yong Sung Kim et al: "Detection of genetic alterations in the human gastric cancer cell lines by two-dimensional analysis of genomic DNA", pages 297-308  --	1-20
P,X	PNAS, Volume 99, No. 16, 2002, Jingfeng Li et al: "Not subtraction and Not-specific microarrays to detect copy number and methylation changes in whole genomes", pages 10724-10729  -- -----	1-20



# INTERNATIONAL SEARCH REPORT

International application No.  
PCT/SE02/00788

## Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:  
because they relate to subject matter not required to be searched by this Authority, namely:
  
2. ☒ Claims Nos.: 14-17  
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:  
  
see next sheet
  
3. ☐ Claims Nos.:  
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

## Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
  
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest : ☐ The additional search fees were accompanied by the applicant's protest.  
☐ No protest accompanied the payment of additional search fees.

Claims 14-15 relate to fragments obtained by the methods described in any of the proceeding claims (1-13). Claims 16-17 relate to representations comprising the nucleic acids obtained by the methods described in any of claims 1-13. These claims relate to an extremely large number of possible fragments and representations. The claims include selected fragments from the digestion of any possible genomic material by any biochemical or chemical approach. The fact that the fragments and representations should be obtained by the method described in any of the proceeding claims is not a limiting feature of the fragments or representations.

Support within the meaning of Article 6 PCT and / or disclosure within the meaning of Article 5 PCT is to be found for only a very small proportion of the fragments and representations claimed. In the present case, the claims so lack support, and the application so lacks disclosure, that a meaningful search over the whole of the claimed scope is impossible. No specific search for these claims has been carried out.

Even those parts of the claims that relate to fragments and representations obtained by digesting nucleic acid with the restriction endonuclease *NotI*, include an extremely large number of fragments not limited by any feature given them by the method by which they are prepared. The claims include previously known fragments and representations, such as the *NotI* linking clones shown in *Nucleic acid research*, 2000, vol 28, no 7, p 1635-1639: "NotI clones in the analysis of the human genome", Zabarovsky et al (see especially p 1636 left column paragraph 4).

## INTERNATIONAL SEARCH REPORT

Information on patent family members

06/07/02

International application No.

PCT/SE 02/00788

Patent document cited in search report			Publication date	Patent family member(s)		Publication date
WO	9923256	A1	14/05/99	AU	1449799 A	24/05/99
				CA	2307674 A	14/05/99
				EP	1032705 A	06/09/00
				JP	2001521754 T	13/11/01
-----						
WO	9842871	A1	01/10/98	EP	0920536 A	09/06/99
				JP	2000516104 T	05/12/00
				US	5958738 A	28/09/99
				US	6235503 B	22/05/01
-----						